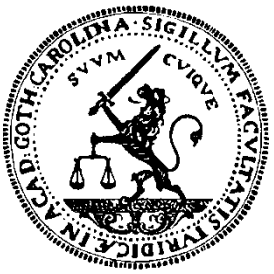

FACULTY OF LAW

LUND UNIVERSITY

Le Thao Vi Nguyen

AI Memorisation and the Reproduction Right:

Protected Works in Generative AI Models under EU Copyright Law

JAEM03 Master Thesis

European Business Law

30 higher education credits

Supervisor: Aurelija Lukoseviciene

Term of graduation: Spring 2026

Contents

Preface

My two-year journey at Lund University has now reached its final chapter, and yet everything still feels as if it happened only yesterday. As I write these final pages, I find myself thinking back to the first letter I sent to Lund, written with hope and curiosity about learning, from one of the world's most influential legal orders, the values that guide legal regulation in a changing world. This thesis, focusing on the relationship between copyright and generative AI, brings me back to some of the reasons that first brought me here. It touches upon questions that have long stayed with me: how law responds to change, how value and power are distributed, and how legal rules can remain faithful to the human beings and fundamental values they are meant to serve.

Reaching this point has not been a journey I have taken alone. I would like to express my sincere appreciation to Aurelija Lukoseviciene, my supervisor, for her guidance, patience and support throughout the process of writing this thesis. I was also fortunate to have learned intellectual property law, especially copyright law, from her during my studies at Lund. Her teaching did not only give me knowledge of the subject, but also strengthened my interest in it. I am also thankful to all the dedicated teachers at Lund University who have shaped my academic journey over these two years. Their knowledge, generosity and commitment to teaching have made my time here not only intellectually demanding, but also deeply meaningful.

My thanks also go to my classmates, friends, and all the kind people I have been fortunate to meet in Lund. I am grateful for all the paths that crossed mine during this journey.

Finally, I owe my deepest gratitude to my family and friends in my homeland which is half a world away, who have always sent me their love, trust and support. Their unwavering belief in my decisions and ambitions has been the enduring foundation beneath every step that has brought me to this moment.

Lund, May 2026

Nguyen, Le Thao Vi

Summary

Generative AI has intensified existing copyright debates by raising questions not only about the use of protected works as training data or the generation of infringing outputs, but also about what may be retained inside trained models themselves. This thesis focuses on memorisation in Generative AI models, particularly large language models, and examines whether such memorisation can amount to reproduction of protected works under EU copyright law.

The analysis begins by explaining the technical background of LLMs and the phenomenon of memorisation. It distinguishes memorisation from general learning, statistical abstraction, extraction, regurgitation and ordinary output similarity. While AI models are designed to generalise from large datasets, technical studies show that they may, under certain conditions, retain and reproduce specific training examples, including near-verbatim or verbatim parts of protected works. This makes memorisation legally relevant, especially where protected expression can be reconstructed or extracted from the model with a sufficient degree of reliability.

EU copyright law gives the reproduction right a broad, functional and technologically neutral scope. A copy does not need to be directly perceptible by humans, nor does it need to take the same form as the original work. However, the copied subject matter must still contain protected expression and be identifiable with sufficient precision and objectivity. On this basis, the thesis argues that memorisation may amount to reproduction where the protected expression of a specific work is retained at model level and can be reconstructed, extracted or regurgitated with sufficient reliability. At the same time, this conclusion must remain limited. It does not mean that an entire model, dataset, embedding, token or internal parameter should automatically be treated as a copy. The legal assessment must focus on whether identifiable protected expression has been retained in a form that performs a copy-like function.

The thesis then considers whether existing EU copyright exceptions can justify such reproduction. The exceptions for temporary copying, incidental inclusion and text and data mining under Article 4 of the DSM Directive provide only limited answers. In

particular, Article 4 may justify certain technical copies made for automated analysis, but it does not comfortably cover the continued retention of protected expression in a trained model where that expression remains capable of later reproduction.

The final part of the thesis evaluates possible responses to memorisation-related copyright risks. Output filters may reduce visible infringement, but they cannot fully address the problem if protected expression has already been retained inside the model. A more coherent approach should combine clearer legal standards, effective transparency obligations, workable licensing mechanisms, fair remuneration for rightholders and proportionate technical mitigation measures. The thesis therefore concludes that AI memorisation can, in specific circumstances, fall within the scope of the reproduction right under EU copyright law, while also showing that the current legal framework still leaves significant uncertainty for rightholders, AI providers and future regulation.

Abbreviations

AI	Artificial Intelligence
AI Act	Regulation (EU) 2024/1689 laying down harmonised rules on Artificial Intelligence
Berne Convention	Berne Convention for the Protection of Literary and Artistic Works
CJEU	Court of Justice of the European Union
DSM Directive	Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market
EU	European Union
EUIPO	European Union Intellectual Property Office
GenAI	Generative Artificial Intelligence
InfoSoc	Directive (EU) 2001/29/EC on copyright in the information society
LLM(s)	Large Language Model(s)
TDM	Text and Data Mining
WCT	The WIPO Copyright Treaty
WPPT	The WIPO Performances and Phonograms Treaty

1 Introduction

1.1 Background

Since the release of ChatGPT on 30 November 2022, Generative AI (GenAI) has moved from a technology mainly limited to the research community to a global social phenomenon. Within a very short period, GenAI systems, especially AI systems based on large language models (LLMs), have been integrated into many aspects of education, intellectual work, communication, services, and content creation. This rapid popularity has also triggered legal and social criticism over the fact that AI systems are built on massive amounts of data which contains a large number of copyrighted works.

Debates and litigation concerning GenAI and copyright have mainly focused on more visible issues, such as the scraping of copyrighted data, the use of copyrighted works to train AI without authorisation, GenAI outputs that are too similar to or reproduce existing works verbatim, and the copyright status of AI-generated outputs. Usually, claims against AI centred on two ends of the process, either at the input stage, where data is collected for AI training, or at the output stage, where outputs are generated in response to user interactions.

The years 2024-2025 saw a significant increase in lawsuits against AI companies.[1] While the familiar issues mentioned above remain common litigation strategies used by plaintiffs, another argument has also gradually appeared more often in the claims, which requires courts to address it directly: arguments concerning the phenomenon of 'memorisation' in trained AI models, an issue located inside the model. [2] This 'memorisation' phenomenon generally refers to a model's retention of specific information from its training data, commonly associated with the possibility that such information may later be recalled and revealed through outputs generated in response to certain prompts.

[1] 'Master List of lawsuits v. AI, ChatGPT, OpenAI, Microsoft, Meta, Midjourney & other AI cos.' (*ChatGPT Is Eating the World*, 26 August 2024) <https://perma.cc/QZ4B-5WX9>.

[2] Megan K Bannigan and others, 'AI Intellectual Property Disputes: The Year in Review and What's Ahead' (*Debevoise & Plimpton*, 2 December 2025) <https://perma.cc/S392-3Q3H>.

From a technical perspective, the capability of AI models to memorise input data is no longer a speculative concern. As early as 2017, the first research works began to provide evidence and warnings about memorisation. [3] Up to the present, an increasing number of technical studies have been developed to show that LLMs can memorise, retain, and, under certain conditions, be exploited to recover training data, including in production models.

This phenomenon has changed the way GenAI models are understood. The concern is no longer confined to the similarity between a particular output and an existing work, but extends to the trained model itself and to the internal conditions that make such output possible. In other words, memorisation raises the possibility that the protected expression of a copyrighted work may be retained within the AI model, or in a technical form functionally equivalent to a copy, even before any output is generated. At this point, the legal focus is not only whether the final output infringes copyright, but also whether the formation, existence, and operation of the model itself may constitute unauthorised reproduction and therefore a copyright infringement.

In US litigation, the issue of AI model memorisation itself constituting copyright infringement has been clearly set out in the claims of The New York Times,[4] as well as mentioned in the output-based claims of authors [5]. Developments in Europe show that, although the number of cases remains limited, debates on how memorisation in GenAI models triggers copyright are also attracting academic attention.

In the literature, there is no longer much dispute over the possibility that AI models can memorise and reproduce training data. However, there are opposing approaches on how this phenomenon could be treated, and whether the parameters inside AI models could

[3] eg Chiyuan Zhang and others, 'Understanding Deep Learning Requires Rethinking Generalization' (International Conference on Learning Representations ('ICLR'), Toulon, April 2017); Satrajit Chatterjee, 'Learning and Memorization' (2018) 80 PMLR 755.

[4] *The New York Times Co v Microsoft Corp* No 1:23-cv-11195 (SDNY 27 December 2023) ('*The NYT v Microsoft*'); *The New York Times Co v Perplexity AI Inc* No 1:25-cv-10106 (SDNY 21 October 2024) ('*The NYT v Perplexity*').

[5] *Kadrey v Meta Platforms Inc* No 3:23-cv-03417 (ND Cal 7 July 2023); *Authors Guild v OpenAI Inc* No 1:23-cv-08292 (SDNY 19 September 2023); *Bartz v Anthropic PBC* No 3:24-cv-05417 (ND Cal 19 August 2024); *Re OpenAI Inc Copyright Infringement Litigation* 1:25-md-03143 (SDNY 11 April 2025); *Denial v OpenAI Inc* No 3:25-cv-05495 (ND Cal 24 July 2025); *Carreyrou v Anthropic PBC* No 3:25-cv-10897 (ND Cal 22 December 2025).

be regarded as protected expressions in the legal sense. This contrast has been clearly reflected in the different decisions of two courts in two European countries. In *Getty Images v Stability AI* in the UK, the High Court ruled in favour of the AI company by rejecting the view that model weights should be treated as copies. [6] By contrast, in *GEMA v OpenAI* in Germany, the Munich Regional court ruled that OpenAI's model memorisation constituted an infringement of the reproduction right and gave rise to legal liability. [7]

At present, three practical issues remain. First, most litigation concerning GenAI in general is still taking place in the United States, but there has not yet been a final ruling on the issue of memorisation. EU-level case law does not yet exist, and guidance or EU positions on this issue remain unclear. Second, many important cases are still at their early stage or have only involved rulings on pleadings, discovery, or preliminary theories, without a full merits judgment, while academic debates remain active and contain many opposing views. Third, the science of memorisation itself is still developing very quickly. New measurement methods and studies from 2025-2026 suggest that the level of extractability may be higher than previously understood, while also highlighting differences between models, types of data, and the techniques used.

1.2 Research Questions

The broad and 'technological neutrality' nature [8] of the concept of reproduction makes it especially important to determine how far the scope of this concept may extend in the context of new technologies. Against this background, this thesis focuses on memorisation in GenAI models, from the perspective of analysing the reproduction right under EU copyright law.

[6] *Getty Images (US) Inc v Stability AI Ltd* [2025] EWHC 2863, [2026] Bus LR 582, [2025] WLR(D) 571.
[7] *GEMA v OpenAI* No 42 O 14139/24 (LG München I 11 November 2025).
[8] 'As it is commonly understood, the principle of technological neutrality prescribes that laws can and should be developed in such a way that they are independent of any particular technology, neither favoring nor discriminating against specific technologies as they emerge and evolve.'. Carys J Craig, 'Technological Neutrality: Recalibrating Copyright in the Information Age' (2016) 17(2) Theoretical Inquiries in Law 601, 604-605.

The main research question of this thesis is: *How does memorisation in Generative AI models challenge the concept of reproduction under EU copyright law, and to what extent can the existing EU copyright framework address this problem?*

To clarify this main question, the thesis develops four sub-questions:

> *First*, how can memorisation in Generative AI models be understood and evidenced from a technical perspective?
>
> *Second*, how could EU copyright law legally characterise the technical manifestations of memorisation in GenAI models when determining whether protected expression has been reproduced?
>
> *Third*, if memorisation is characterised as reproduction, to what extent can existing EU copyright exceptions and limitations, particularly the temporary copying exception, incidental inclusion and text and data mining exceptions, apply to AI memorisation issues?
>
> *Fourth*, how have the legal risks raised by AI memorisation been addressed through existing technical and legal approaches, and what possible future responses may be considered?

1.3 Methodology and Delimitations

1.3.1 Methodology and Materials

This thesis primarily adopts a legal dogmatic, or doctrinal legal, method. This method can be understood as a structured analysis of legal rules, principles and concepts within an existing legal framework, aimed at interpreting the current law, clarifying ambiguities and addressing legal gaps. [9]

[9] J M Smits, 'What Is Legal Doctrine? On the Aims and Methods of Legal-Dogmatic Research' in R van Gestel, H Micklitz and E L Rubin (eds), *Rethinking Legal Scholarship: A Transatlantic Dialogue* (Cambridge University Press 2017) 207-228.

In this thesis, the analysis primarily concerns the interpretation and application of existing EU copyright law. It examines whether memorisation in GenAI models may fall within the scope of the reproduction right and whether existing exceptions and limitations may justify the act or limit the liability of AI developers. It seeks to clarify how the relevant principles, rules and concepts interact, and how they may apply to the technically new phenomenon of AI memorisation.

Beyond describing the existing law, the thesis also considers how this legal framework may develop in response to the issue. On the basis of the doctrinal analysis, it explores possible future directions, including legal solutions, policy choices and regulatory responses. This exploratory part aims to assess which approaches may fit more coherently within the structure and rationale of EU copyright law, rather than to propose a definitive legislative solution.

The legal analysis is based mainly on EU copyright instruments, particularly the InfoSoc Directive and the DSM Directive, together with relevant case law of the Court of Justice of the European Union (CJEU). These sources are used to analyse the scope of the reproduction right, the requirements of a copy, and the operation of exceptions and limitations. Academic commentary is used to identify different doctrinal positions.

Where relevant, the thesis also refers to international copyright instruments and historical materials, including the Berne Convention, the WIPO Copyright Treaty and discussions concerning copyright protection in the digital environment. These materials are treated as contextual materials that help explain the historical and international background that has influenced, to some extent, the formation and interpretation of the reproduction right in EU copyright law.

Because memorisation is also a technical phenomenon, the thesis relies on selected technical literature on machine learning and GenAI. These literatures are used to explain the training process of LLMs, AI memorisation and methods used to detect or evidence memorisation. The thesis does not conduct an independent empirical or technical assessment of any AI model. Those technical literature is used only to provide the factual and conceptual background necessary for the legal analysis.

The thesis also uses selected national and foreign case law, litigation materials and policy documents where they are directly relevant to the legal framing of AI memorisation. For instance, *GEMA v OpenAI* is used as a European case study because it directly concerns model memorisation and the reproduction right. Other materials, including *Getty Images v Stability AI*, selected US litigation, European Parliament materials and expert reports, are used to illustrate emerging arguments and possible regulatory responses. They are not treated as binding sources of EU law, and the thesis does not undertake a full comparative analysis of German, UK or US copyright law.

1.3.2 Delimitation

This thesis is limited to EU copyright law. It does not provide a full comparative analysis of national copyright laws, although selected national and foreign materials are used where they are useful for illustration or contextual analysis.

This thesis discusses GenAI in general, but its main focus is on LLMs and text-based memorisation. This focus is chosen because LLM-based systems are widely used in practice [10] and are often discussed in technical and legal debates on memorisation. The analysis mainly concerns protected literary works, especially natural language texts. Examples from image, music, code or other GenAI systems are used only when they help explain the legal or practical issues.

AI training data may include copyrighted and non-copyrighted materials, protected and unprotected subject matter, facts, ideas, public domain works and licensed works. This thesis does not analyse all types of training data. It focuses on situations where the material allegedly memorised or reproduced is assumed to be a protected work, and where the relevant output or extracted content, if any, is identical or substantially similar to that work. Questions of originality, ownership, access and similarity are therefore not examined in detail, except where they are necessary to frame the analysis of memorisation and the reproduction right.

[10] Nestor Maslej and others, 'The AI Index 2025 Annual Report' (*Stanford University, Institute for Human-Centered AI*, April 2025) <https://perma.cc/RZN3-H9V2>.

The terms ‘memorisation’ or ‘memory’ may be used in some writings in a broad sense, referring generally to the way a model internalises information from training data as part of the learning process. [11] However, in current technical and legal debates on AI, ‘memorisation’ is often understood separately from generalisation and in a narrower sense: the phenomenon where a model retains specific training examples in a way that allows them to be reproduced, regurgitated or extracted under certain conditions. This thesis uses ‘memorisation’ in this narrower sense, in line with its research focus and the common use of the term in this context.

GPAI tools, specifically ChatGPT and NotebookLM, were used only for translation, grammar correction and limited source-search assistance; no AI-generated content has been used. The thesis remains the author’s own work and responsibility.

1.4 Outline

To address the research questions in a clear and coherent manner, this thesis is structured into six chapters. Chapter 1 introduces the background of the research, the research question, methodology, delimitations and overall structure of the thesis.

Chapter 2 provides the technical foundation for the legal analysis. It explains how Generative AI models, particularly LLMs, are trained and operated through tokenisation, embeddings, parameter optimisation and output generation. It then examines the phenomenon of memorisation, its main forms, and the technical methods used to detect or evidence memorisation.

Chapter 3 analyses the reproduction right in the context of Generative AI. It first explains the scope of the reproduction right under EU copyright law, especially its broad and technologically neutral character. It then examines whether AI memorisation may satisfy the legal requirements of reproduction, including protected expression,

[11] eg Ivo Emanuilov and Thomas Margoni, ‘Memorisation in Generative Models and EU Copyright Law: An Interdisciplinary View’ (*Kluwer Copyright Blog*, 26 March 2024) <https://legalblogs.wolterskluwer.com/copyright-blog/memorisation-in-generative-models-and-eu-copyright-law-an-interdisciplinary-view/> accessed 24 May 2026; Madhur Panwar and others, 'For Better or for Worse, Transformers Seek Patterns for Memorization' (*NeurIPS*, 3 December 2025) <https://neurips.cc/virtual/2025/loc/san-diego/poster/119570> accessed 24 May 2026.

fixation, identifiability and recognisability. This chapter also discusses the main arguments for and against treating model-level memorisation as reproduction.

Chapter 4 considers whether existing exceptions and limitations under EU copyright law may apply or limit liability where AI memorisation is treated as reproduction. It focuses on the exception for temporary acts of reproduction under Article 5(1) of the InfoSoc Directive, incidental inclusion under Article 5(3)(i) of the InfoSoc Directive, and the text and data mining exception under Article 4 of the DSM Directive. The chapter also considers the three-step test as a final limit on the interpretation and application of copyright exceptions.

Chapter 5 discusses possible responses to AI memorisation. It briefly notes technical measures to reduce memorisation, but focuses mainly on broader legal and policy responses and the allocation of obligations among AI providers.

Chapter 6 concludes the thesis by answering the research question and summarising the main findings. It argues that AI memorisation may amount to reproduction. It also reflects on the broader implications of this conclusion for rightholders, AI providers, licensing practices and the future development of EU copyright law.

2 Memorisation of Generative AI Model

2.1 Generative AI Models and the Memorisation Phenomenon

2.1.1 Technical Foundations of Generative AI

To analyse the phenomenon of memorisation in AI, it is first necessary to understand how such models are constructed and operate. At a fundamental level, an AI model is a mathematical system trained on large datasets to learn how to predict or generate outputs from given inputs. In the case of LLMs, the core task is typically next-token prediction within a sequence of text.

The learning process can be broadly divided into three main stages:

(1) Input processing

Input text is not processed as complete words but is segmented into smaller units called tokens[12] (which may correspond to words, subwords, or characters). Each token is then mapped into a numerical vector within a high-dimensional space through an embedding mechanism.[13] These vectors provide the machine-readable representations through which the model can analyse relationships between tokens and generate predictions.[14]

(2) Parameter optimisation through training

Using the token vectors generated at the previous stage as inputs, the model iteratively adjusts its parameters, including embedding weights and transformer weights,[15] in order to minimise prediction error (loss). Through repeated exposure to training data, the model encodes statistical patterns of the training data into billions of parameters. However, this mechanism also enables the model, in certain cases, to retain specific sequences from the training data, particularly when such sequences are repetitive or distinctive.[16]

(3) Generation

Given an input prompt, the model generates output by iteratively predicting the next token based on learned probability distributions. Each newly generated token becomes

[12] Tokens are the discrete textual units processed by language models. Depending on the tokenizer, a token may correspond to a whole word, part of a word, punctuation mark, or character. Tokenization refers to the process of converting raw text into such units before numerical encoding. See Daniel Jurafsky and James H Martin, 'Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models' (2026) 3rd draft edn, Online manuscript released 6 January 2026 <https://web.stanford.edu/~jurafsky/slp3> accessed 22 May 2026, ch 2.

[13] See Jurafsky and Martin (n 12) ch 5; EUIPO, *The Development of Generative Artificial Intelligence from a Copyright Perspective* (EUIPO, 2025) 147.

[14] Jurafsky and J Martin (n 12) ch 3.

[15] In neural networks, model weights or biases are learned parameters of the model. They are typically initialised before training and then iteratively updated as the model is exposed to training data, so that the model better satisfies the optimisation objective specified by its designers. In transformer models, the attention mechanism does not mean that each item of training data is stored in a discrete location within the model. Rather, it is a computational mechanism which allows the model, for a given input context, to determine the relative relevance of tokens or elements within the sequence. Thus, training data shapes the model by influencing its learned parameters, whereas attention weights are dynamically computed during the processing of a particular input and indicate which parts of that input have greater influence on the resulting output. See *Getty* (n 6) [5]; *GEMA* (n 7) [14]; for a more technical understanding, see also Jurafsky and Martin (n 12) ch 6-8.

[16] Nicholas Carlini and others, 'Quantifying Memorization Across Neural Language Models' (ICLR, Rwanda, 2023).

part of the context for subsequent predictions. This sequential mechanism may, in certain circumstances, result in the reproduction of sequences that appeared in the training data, where their probability remains sufficiently high.

Although these stages are analytically distinguishable, they are not fully separable in technical operation. Modern GenAI systems function through densely interconnected processes in which data transformation, parameter adjustment, and output generation continuously interact and feed back into one another.[17]

An important feature of modern AI models is that the way they process and store information inside the model cannot be directly and fully observed. Although there are many technical methods for inferring, measuring or testing how GenAI operates, it remains complex to determine how content is retained in the model, where it is retained, and from what point in the training process.

This technical opacity forms the basis for the later legal debate on whether the fact that a work is introduced into AI models, converted into tokens, represented as embeddings, and involved in the adjustment of model parameters, as described in stages one and two above, can be sufficient to be regarded as reproduction within the meaning of copyright law.

In this context, the stage of output generation plays a particularly important evidential role. In practice, outputs that are identical or highly similar to copyrighted works are the evidence most strongly relied on by plaintiffs in AI litigation to hold AI companies liable. Alongside such output-based evidence, claimants also use the output as a basis for arguing that the model must have received, retained and become capable of reproducing the protected content internally. [18]

[17] Feder Cooper and James Grimmelmann, 'The Files are in the Computer: On Copyright, Memorization, and Generative AI' (2025) 100(1) Chi-Kent L Rev 141, 153-155.

[18] In many cases, memorisation is invoked only as factual or evidentiary support, for example to show access, copying, output infringement, or harm, as in *Authors Guild* (n 5); *Re OpenAI* (n 5); *Kadrey* (n 5). In other cases, memorisation is pleaded as a direct basis of liability: the claimant alleges that the trained model itself infringes copyright because it contains or stores copies of protected works, as in *The NYT v Microsoft* (n 4), *Getty* (n 6), and *GEMA* (n 7).

2.1.2 Empirical Emergence of Memorisation Concerns

In the machine learning literature, memorisation may be defined as a model learning a direct association between features of the input and specific outputs, rather than generalising from data or applying a general method to infer the result. [19] More specifically, 'a model has "memorized" a piece of training data when (1) it is possible to reconstruct from the model (2) a (near-)exact copy of (3) a substantial portion of (4) that specific piece of training data.'.[20]

Since the early stages of machine learning and AI technology, many experts have studied and warned that deep networks have the capacity to memorise data, including random labels or noise. [21] Concerns regarding the ability of AI systems to reproduce training data have emerged progressively, gaining significant attention with the rapid development of GenAI since 2023. Observers began to note that generative models occasionally produce outputs that appear excessively similar to identifiable source materials. For instance, a frequently cited practical example is the *Harry Potter* case, where experimental evidence has shown that when GPT-3 is prompted with the opening of a chapter from *Harry Potter and the Philosopher's Stone* [22], the model may continue by reproducing a substantial portion of the original text almost verbatim before deviating.

In the domain of code generation, studies have shown that LLMs can reproduce training data to a significant extent. For example, experiments on the CodeGen-Mono-16B model reported the successful extraction of approximately 47% of targeted sequences

[19] Tim Hartill and others, 'Do Smaller Language Models Answer Contextualised Questions Through Memorisation Or Generalisation?' (2023) <https://doi.org/10.48550/arXiv.2311.12337>.
[20] Cooper and Grimmelmann (n 17) 142.
[21] eg Zhang (n 3); Chatterjee (n 3); Aparna Elangovan, Jiayuan He, and Karin Verspoor 'Memorization vs. Generalization: Quantifying Data Leakage in NLP Performance Evaluation' (2021) <https://doi.org/10.48550/arXiv.2102.01818>; Nicholas Carlini and others, 'The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks' in *Proceedings of the 28th USENIX Security Symposium* (USENIX Association 2019) 267.
[22] eg Javier Rando and Florian Tramèr 'Extracting (Even More) Training Data From Production Language Models' (*SPYLab*, 2 July 2024) <https://spylab.ai/blog/training-data-extraction/> accessed 22 May 2026; A. Feder Cooper and others, 'Extracting memorized pieces of (copyrighted) books from open-weight language models' (2025) <https://doi.org/10.48550/arXiv.2505.12546>; Timothy B. Lee, 'Meta's Llama 3.1 can recall 42 percent of the first Harry Potter book' (*Understanding AI*, 12 Jun 2025) <https://www.understandingai.org/p/metas-llama-31-can-recall-42-percent> accessed 22 May 2026; Ahmed Ahmed and others, 'Extracting books from production language models' (2026) arXiv:2601.02671 <https://doi.org/10.48550/arXiv.2601.02671>.

through data extraction attacks. Notably, the extracted data included sensitive information such as names, email addresses, usernames, and API keys.[23] Similarly, GitHub Copilot has been observed reproducing verbatim segments of well-known code, e.g, code from Quake III and from individual programmers, including both functional code and non-functional comments, suggesting direct memorisation rather than mere pattern learning.[24] GitHub, while denying that its models reproduce or store training data[25], nonetheless acknowledges Copilot's 'recitation capabilities' [26] and accordingly provides code referencing and filtering mechanisms to detect and limit outputs that 'match' or are 'near matches' to existing public code.[27]

These observations have raised concerns that such models may not merely learn general patterns, but, in some cases, memorise and reproduce elements of their training data.
In response to allegations and inferences concerning memorisation, technology companies, understandably, tend to deny that the AI models they own are capable of memorisation. [28] The usual counter-argument is that the model only learns patterns. If any memorisation occurs during training, the 'remembering' of specific examples functions only as an intermediate step for optimisation in the early stage of training. Later, when the model learns more general rules, verbatim memorised traces are replaced by more abstract representations, and the final model therefore only 'understands' in order to generalise, rather than retaining training data inside its machinery. [29] Some commentators in the technical field also emphasise that GenAI

[23] Ali Al-Kaswan, Maliheh Izadi, and Arie van Deursen, 'Traces of Memorisation in Large Language Models for Code' in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (ACM 2024).
[24] *Doe 1 v GitHub Inc* No 4:2022cv06823 (ND Cal 22 January 2024); Albert Ziegler, 'GitHub Copilot research recitation' (*Github Blog*, 30 June 2021) <https://github.blog/ai-and-ml/github-copilot/github-copilot-research-recitation> accessed 22 May 2026.
[25] In the General section on its official website, GitHub denies that Copilot engages in copy-and-paste, emphasising that 'GitHub Copilot generates suggestions using probabilistic determination.'. See GitHub, 'Frequently asked questions: General: Does GitHub Copilot "copy/paste"?' (*GitHub*) <https://github.com/features/copilot> accessed 22 May 2026.
[26] Ivo Emanuilov and Thomas Margoni, 'Forget Me Not: Memorisation in Generative Sequence Models Trained on Open Source Licensed Code' (SSRN, 2024) <http://dx.doi.org/10.2139/ssrn.4720990>.
[27] GitHub, 'Finding Matching Code' (*GitHub Docs*) <https://docs.github.com/en/copilot/how-tos/get-code-suggestions/find-matching-code> accessed 24 May 2026.
[28] Alex Reisner, 'AI's Memorization Crisis' (*The Atlantic*, 9 January 2026) <https://www.theatlantic.com/technology/2026/01/ai-memorization-research/685552/> accessed 24 May 2026.
[29] This phenomenon is also sometimes referred to as 'grokking'. See also Bahadir Akdemir, 'Meta, Google & NVIDIA's Landmark Study: When LLMs Stop Memorizing and Starts Generalizing' (*Medium*, 11 June 20025) <https://medium.com/%40akdemir_bahadir/how-much-do-language-

systems are designed to generalise, rather than to memorise specific data. [30] Nevertheless, due to model scale, data structure, and optimisation processes, the boundary between generalisation and memorisation is not clearly defined.

2.1.3 AI Memorisation Categories

In machine learning research, memorisation is not a unitary phenomenon but may be classified according to its underlying mechanisms and functional role.

From a design perspective, a distinction can be drawn between intentional memorisation and unintentional memorisation. Intentional memorisation occurs in models explicitly designed to retain training data as part of their functioning (e.g. KNN or SVM). However, such models are not the focus of this thesis. Instead, the analysis concentrates on GenAI models, where memorisation is generally regarded as an unintended by-product, or even a 'bug', although in some cases it is also recognised as a necessary feature that may contribute to generalisation.[31]

The first phenomenon commonly discussed in relation to memorisation is overfitting, where the model learns training data too closely and fails to generalise to unseen inputs. Overfitting typically arises from limited or noisy datasets, or excessive training relative to data diversity.[32]

Other phenomena, often referred to as unintended or eidetic memorisation,[33] also exists, is not attributable to training error but rather to intrinsic properties of large-scale models. In such cases, the model may reproduce rare or distinctive sequences from training data, even when these are not relevant to the intended task. Importantly, memorisation is not

models-actually-remember-a-deep-dive-into-ai-memory-and-capacity-f7a71e55a93e> accessed 24 May 2026.

[30] Leo Gao and others, 'The Pile: An 800GB Dataset of Diverse Text for Language Modeling' (2020) arXiv:2101.00027 <https://doi.org/10.48550/arXiv.2101.00027>; Delip Rao, 'The NYT vs. OpenAI Copyright Case: A Technical Introduction - Part I' (*AI Research & Strategy*, 8 January 2024) <https://deliprao.substack.com/p/the-nyt-vs-openai-copyright-case> accessed 24 May 2026; Lee Gesmer, 'Copyright and the Challenge of Large Language Models (Part 3)' (*Mass Law Blog*, 27 November 2024) <https://masslawblog.com/copyright/copyright-and-the-challenge-of-large-language-models-part-3/> accessed 24 May 2026.

[31] Emanuilov and Margoni (n 26) 10.

[32] Nicholas Carlini and others, 'Extracting Training Data from Large Language Models' in *Proceedings of the 30th USENIX Security Symposium* (USENIX Association 2021) 2633.

[33] Carlini and others (n 21); Carlini and others (n 32) 2635-2642.

necessarily detrimental. In long-tailed data distributions[34], the memorisation of rare instances may contribute to improved generalisation performance.[35]

The degree of memorisation is influenced by several key factors. First, model scale, which shows that larger models exhibit greater memorisation capacity. Second, sequences that occur repeatedly in training data are more likely to be memorised. Third, memorised content may only be revealed when the model is prompted with sufficiently specific or extended input.[36]

2.2 Detecting AI Memorisation

The memorisation phenomenon in AI models, particularly GenAI systems, is supported not only by theoretical considerations but also by extensive empirical research employing rigorous experimental methodologies.

Among the earliest and most influential studies are those of Carlini and colleagues, whose successive works helped establish memorisation as a serious research topic rather than a speculative concern. Their studies developed quantitative methods for assessing unintended information retention and demonstrated that, under certain conditions, generative models may reproduce memorised training content or leak sensitive sequences. [37]

During the period from 2024 to 2026, research on AI memorisation entered a more mature stage. The research community developed a diverse ecosystem of methods for measuring the degree, forms, and potential legal implications of memorisation. These methods differ not only in technical design, but more importantly in the distinct aspects of memorisation they seek to measure, ranging from verbatim reproduction,[38] near

[34] For 'long-tailed distribution' definition, see Chongsheng Zhang and others, 'A Systematic Review on Long-Tailed Learning' (2024) ArXiv:2408.00483 <https://doi.org/10.48550/arXiv.2408.00483>.
[35] Vitaly Feldman, 'Does learning require memorization? A short tale about a long tail' in Konstantin Makarychev and others (eds), *STOC '20 Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* (ACM, 2020) 954.
[36] Emanuilov and Margoni (n 26) 12.
[37] eg, for early publicly available versions: Carlini and others (n 21) in year 2018; Carlini and others (n 32) in year 2021; Nicholas Carlini and others, 'Extracting Training Data from Diffusion Models' in *Proceedings of the 32nd USENIX Conference on Security Symposium* (USENIX Association, 2023) in year 2023 and so on.
[38] Carlini and others (n 32); Al-Kaswan, Izadi and van Deursen (n 23).

reproduction,[39] and identification of training data,[40] to situations in which a model may 'remember but not reveal'.[41]

One common approach, descriptively referred to as prompt-based extraction [42], uses different forms of prompts to trigger and observe model responses. Some studies also combine this with partial access to input information or intervention in the training data (by inserting unique or rare sequences, so-called 'canaries' or 'secrets', into training datasets[43]). These techniques rely on testing large numbers of prompts and then comparing outputs with the original works or with external reference datasets, thereby assessing both the extent and stability of memorisation across different models.

A common form of prompt-based extraction divides the training data into a prefix and a suffix. The model is provided with the prefix and tasked with predicting the continuation, and correct reconstruction of the suffix may indicate memorisation of the original sequence. [44] This approach underlies various forms of targeted data extraction attacks, where partial knowledge is used to recover the remainder of a sequence.[45] Within this framework, 'exact match' [46] (identical reproduction) refers to cases where the model reproduces a sequence identically, character by character. In addition, another phenomenon, referred to as 'fuzzy matching'[47], uses measurement methods to identify outputs that are not identical but remain substantially similar to the source material, thereby avoiding the omission of non-verbatim reproductions that may still constitute a substantial part of the original works.

---

[39] Felix B Mueller and others, 'LLMs and Memorization: On Quality and Specificity of Copyright Compliance' in Sanmay Das and others (eds), *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society (AIES-24)* (AAAI Press, 2024) vol 7, no 1.

[40] Carlini and others (n 21).

[41] Al-Kaswan, Izadi, and van Deursen (n 23).

[42] An umbrella term covering prefix-prompting, targeted extraction attacks, adversarial prompting, and iterative continuation or having similar function (if any) methods. Although each paper uses its own terminology (e.g., 'data extraction attacks', 'prefix-continuation', 'jailbreak prompting'), they all share the same underlying mechanism: providing crafted prompts to elicit memorized training data from the model.

[43] Carlini and others (n 32).

[44] A. Feder Cooper and others, 'Extracting memorized pieces of (copyrighted) books from open-weight language models' (SSRN, 2025) Stanford Public Law Working Paper, WVU College of Law Research Paper No 2025-005 <http://dx.doi.org/10.2139/ssrn.5262084>.

[45] Al-Kaswan, Izadi and van Deursen (n 23).

[46] Carlini and others (n 32) 8.

[47] Mueller and others (n 39); Igor Shilov, Matthieu Meeus and Yves-Alexandre de Montjoye, 'The mosaic memory of large language models' (2026) 17 Nat Commun 2142 <https://doi.org/10.1038/s41467-026-68603-0>.

Another branch of research focuses on latent memorisation through methods such as exposure tests,[48] membership inference attacks,[49] and perplexity based measures.[50] These methods examine whether a model displays an unusual preference for rare sequences, reacts differently to previously seen data, or demonstrates a high degree of familiarity with a specific text. Such approaches suggest that training data may reflect internal storage or retention within the model, even where no corresponding material appears in the output.

Beyond output-based methods, some studies have also sought to measure memorisation through internal model signals or through the training process itself. For example, influence functions and training data attribution methods attempt to trace a prediction or a factual assertion back to the training examples that most strongly influenced it. [51] Research on gradient leakage further shows that, in certain distributed training settings, training data may be reconstructed from gradients.[52] However, these methods usually require access to the model, gradients, checkpoints, or the training pipeline. They therefore primarily serve as technical evidence of the possibility that data may be retained within the internal structure of a model, rather than as tools readily applicable to commercial black box LLM systems.

The most recent study available (at the time of this thesis), by Ahmed Ahmed, A. Feder Cooper, and colleagues, published in early 2026 and conducted on the most advanced models commercially available as of August 2025, reported striking findings that further reinforce the strong memorisation capacity of LLMs. Using a two stage method, (1) an initial probe to test the feasibility of data extraction, combined with a Best of N jailbreak technique, and (2) iterative continuation prompts aimed at reconstructing

[48] Carlini and others (n 21); Carlini and others (n 16).
[49] Carlini and others (n 16).
[50] Daphne Ippolito and others, 'Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy' in Maria C Keet, Hung-Yi Lee and Sina Zarrieß (eds), *Proceedings of the 16th International Natural Language Generation Conference* (Association for Computational Linguistics 2023) DOI:10.18653/v1/2023.inlg-main.3.
[51] Pang Wei Koh and Percy Liang, 'Understanding black-box predictions via influence functions' in Doina Precup and Yee Whye Teh (eds), *Proceedings of the 34th International Conference on Machine Learning* (JMLR.org 2017) vol 70.
[52] Ligeng Zhu, Zhijian Liu, and Song Han, 'Deep Leakage from Gradients' in Hanna M Wallach and others (eds), *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (Curran Associates Inc 2019).

entire books, the study found that, for *Harry Potter*, Claude 3.7 Sonnet was able to reproduce 95.8 per cent, Gemini 2.5 Pro reproduced 76.8 per cent, Grok 3 reproduced 70.3 per cent, and GPT 4.1 reproduced nearly 3,200 words despite repeatedly refusing at the outset. [53]

However, despite these developments and the efforts to show that memorisation is a real, persistent and inherent feature of GenAI models, the extent and nature of memorisation in these models remain contested. The disagreement mainly concerns whether the observed phenomena reflect genuine 'memorisation' of training data, or, as some researchers argue [54] and many technology companies rely on,[55] they are merely accidental, rare, or probabilistic occurrences caused by the prevalence of certain patterns and by large-scale statistical generalisation. Nevertheless, as some authors have observed, in certain cases of reproduction, it is difficult to offer any convincing explanation other than that the model has memorised or retained the original training data. [56]

3 The Reproduction Right in the Context of Generative AI

3.1 Reproduction Right in EU Copyright Law

3.1.1 Technological Neutrality and the Broad Interpretation of Reproduction

From the very beginning of the history of copyright law, the right of 'reproduction' has always been a primary and foundational right granted to authors. [57] Reproduction is a

[53] Ahmed and others (n 22).
[54] Nelson Elhage and others, 'Softmax Linear Units' (*Transformer Circuits Thread*, 27 June 2022) <https://transformer-circuits.pub/2022/solu/index.html> accessed 24 May 2026; Mark Chen and others, 'Evaluating Large Language Models Trained on Code' (2021) Arvxiv:2107.03374, 13 <https://doi.org/10.48550/arXiv.2107.03374>.
[55] See, for discussion of memorisation and industry responses, Reisner (n 28); see also arguments advanced by Google, Perplexity, OpenAI and other AI providers in, eg *The NYT v Microsoft* (n 4); *'The NYT v Perplexity* (n 4); *Authors Guild* (n 5); *GEMA* (n 7); Case C-250/25 *Like Company v Google Ireland Ltd* (pending).
[56] Emanuilov and Margoni (n 26) 24.
[57] Jane C Ginsburg and Edouard Treppoz, *International Copyright Law: U.S. and E.U. Perspectives: Text and Cases* (Edward Elgar 2015) 311-312.

right often regarded as the most important within the copyright system, reflecting the core function that the term 'copyright' itself seeks to protect. [58]

From the 1960s onward, and more clearly during the 1970s, copyright law increasingly had to confront challenges arising from successive waves of technological change. Initially, these included phonograms, broadcasting, and reprographic copying. Later, the process of computerisation, with the emergence of computer programs, electronic databases, and digital storage systems, [59] further demonstrated that works could be stored, transmitted, reproduced, and processed through technical systems in ways that no longer correspond to traditional models or explanations.

A recurring pattern in the history of copyright law is that each technological transformation has required reconsideration of the scope and adequacy of existing legal concepts. In the early stages of harmonisation, particularly from the late 1980s and throughout the 1990s, the response of EU copyright law to challenges created by new technologies took place mainly through sector specific directives addressing separate issues individually. The adoption of the InfoSoc Directive marked a more horizontal effort to harmonise the core economic rights within the digital environment.[60]

The InfoSoc Directive made clear that technological development required adjustment of the existing framework in a manner responsive not only to economic realities[61] but also to the 'specific features of contents of product and services'.[62] Traditional concepts of protected copyright rights were expected to adapt through interpretation and recalibration, rather than to be replaced by entirely new concepts. [63] It can be seen that the concept of reproduction developed under EU copyright law reflects technological neutrality and is capable of operating across changing technological forms. This also resonates with WIPO's guide to the Berne Convention on the concept of reproduction, according to which the scope of reproduction is not only broad enough to cover all

[58] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* EU:C:2025:30, Opinion of AG Trstenjak, point 48.
[59] Mihály Ficsor, *The Law of Copyright and the Internet: The 1996 WIPO Treaties, Their Interpretation and Implementation* (Oxford University Press 2002) 4-10.
[60] Péter Mezei, 'A comprehensive guide to the InfoSoc Directive', (2020) 15(1) JIPLP 70, 70-71.
[61] InfoSoc recital (5).
[62] InfoSoc recital (8).
[63] InfoSoc recital (5).

methods of reproduction currently available to human beings, but is also left open to 'all other processes known or yet to be discovered'. [64]

Therefore, the current debate on whether AI-related processes constitute 'reproduction' should not be viewed as an entirely new legal issue, but rather as the latest stage in the continuing development of the concept of reproduction.

Set out in Article 2 of the InfoSoc Directive, the reproduction right is positioned as an exclusive right of the rightholder. No specific definition of 'reproduction' can be found in official EU instruments or in the Berne Convention. Referring to the definition in the Rome Convention, reproduction is defined as the 'making of a copy or copies of a fixation'. [65] Although this definition has been criticised as not providing a 'truly substantive definition of reproduction',[66] it still reflects the common understanding of reproduction at its most basic level.

From the wording of Article 2 of the InfoSoc Directive, the acts and objects covered by the rightholder's reproduction right are extremely broad. They include the reproduction of a work in a 'direct or indirect', 'temporary or permanent' manner, 'by any means and in any form', 'in whole or in part'. [67] Through its case law, the CJEU has also provided a very broad functional interpretation of this right. Accordingly, the extraction of a sentence of only eleven words may be regarded as reproducing part of a protected work, [68] while temporary copies created in the memory of a decoder, on a screen,[69] or cache copies generated during web browsing [70] may also constitute acts of reproduction. In general, reproduction covers any act of making any copies, regardless of the technological process used.

---

[64] Claude Masouyé, *Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971)* (WIPO 1978) 54.
[65] International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organisations (adopted 26 October 1961, entered into force 18 May 1964) 496 UNTS 43 ('Rome Convention') art 3(e)
[66] Ficsor (n 59) 97.
[67] InfoSoc art 2.
[68] Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* EU:C:2009:465 *('Infopaq I').*
[69] Joined Cases C-403/08 and C-429/08 *Football Association Premier League Ltd and Others v QC Leisure and Others and Karen Murphy v Media Protection Services Ltd* EU:C:2011:631.
[70] Case C-360/13, *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd and Others* EU:C:2014:1195 ('*Meltwater'*).

3.1.2 Elements Constituting Reproduction under EU Copyright Law

Since reproduction is, in essence, the act of making a copy, an act can fall within the scope of the reproduction right only if its result creates an object that copyright law can recognise as a copy of a protected work. [71] From the case law and legal analysis concerning the reproduction right, a number of factors may serve as indicators for determining whether the object produced by a technical process can be recognised as a copy within the meaning of copyright law.

(1) First, the reproduced elements, whether the whole work or only part of it, must contain elements of expression capable of reflecting the author's own intellectual creation.

The CJEU has held that the purpose and meaning of this right should be understood as protecting the author's right to authorise or prohibit reproduction in order to cover 'works', that is, 'intellectual creations'. [72] On this basis, copyright law does not require the copied part of the author's work to be an independent work in itself. At the same time, it also does not accept that any part of a work, merely because it derives from a copyrighted work, is automatically a protected copy. The CJEU's reasoning emphasises the value of the elements of intellectual creation that make up a 'work', and it is these values that the reproduction right seeks to protect, in order to ensure that authors can control and receive appropriate remuneration for the use of their works.

It has been confirmed by the CJEU that there is no different treatment between the reproduction of a whole original work and the reproduction of only part of it, provided that what is regarded as the copy reflects the author's personal imprint through the free and creative selection, arrangement and combination of expressive elements. [73] The interpretation of 'reproduction' is not defined in purely quantitative terms, or by reference to a de minimis threshold. [74] Rather, it depends on whether the reproduced elements are capable of 'conveying to the reader the originality of a publication' and of

[71] Ficsor (n 59) 97.
[72] *Infopaq I* (n 68) [32]-[34].
[73] *Infopaq I* (n 68) [44]-[45], *Football Association Premier League* (n 68) [98], Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH and Others.*, EU:C:2011:798 [89]-[99].
[74] AG Trstenjak (n 58) point 58.

'communicating an element which is, in itself, the expression of the intellectual creation of the author'. [75]

(2) Second, the copy must exist in some form of fixation, which may or may not correspond to the format of the original work, and this condition should be understood broadly.

Advocate General Trstenjak once proposed that '"reproduction" (or "reproduction in part") of a work can be defined as fixation of the work (or part only of a work) in a given information medium.'. [76] The guide to the Berne Convention also states briefly that 'it is simply a matter of fixing the work in some material form.'. [77]

Looking back to the Stockholm Revision Conference, [78] the discussions on Article 9(1) Berne Convention indicate that 'reproduction' was understood as requiring fixation in a material form of the work. [79] This element of fixation also includes a requirement of a certain degree of stability. However, this should not be understood as requiring a static or immutable material state. Rather, reproduction should be understood in a functional sense, meaning that there must be a form sufficiently stable to serve as a basis for the indirect communication of the work or for the making of further copies. [80] This distinguishes reproduction from mere performance or direct communication, where the public perceives the work directly without the mediation of a fixed copy. [81]

The international discussion on copyright in the computer environment, continuing from the 1982 Recommendations to the WCT, shows a consistent understanding that 'the reproduction right, as set out in Article 9 of the Berne Convention, [...] fully apply in the digital environment', [82] and that the fixation of a protected work in digital form in an electronic medium, including the memory of a computer system, constitutes

[75] *Infopaq I* (n 68) [47].
[76] AG Trstenjak (n 58) point 56.
[77] Masouyé (n 64) 54.
[78] The Stockholm Revision Conference was the 1967 diplomatic conference for the revision of the Berne Convention, at which the general reproduction right now found in Article 9 of the Convention was introduced.
[79] WIPO, *Records of the intellectual property conference of Stockholm* (WIPO Publication No 311, 1971) 851-855.
[80] Ficsor (n 59) 94-95, 444-445.
[81] Ficsor (n 59) 444-445.
[82] WIPO Copyright Treaty (adopted 20 December 1996, entered into force 6 March 2002) 2186 UNTS 121 ('WCT') Agreed Statement concerning Article 1(4), 8.

reproduction. [83] As noted in section 3.1.1 above, this view has also been clearly reflected by the CJEU in its case law concerning technological contexts.

(3) Third, the copy does not necessarily have to be directly perceived by humans, but the content fixed in it must be capable of being retrieved or identified through appropriate technical equipment, that is, it must have identifiability.

The requirement that a copy must exist in some form of material fixation does not mean that it must take a form that humans can immediately read, hear, or see.

During earlier international discussions on computer programs, electronic storage, and the WIPO Internet Treaties, one recurring issue was whether works stored in machine-readable formats or electronic memory could fall within copyright scope even though they were not directly perceptible by humans. The agreement that direct human perceptibility is not required laid the foundation for extending the concept of reproduction to machine-readable copies, computer memory fixation, and digital storage. [84]

Accordingly, invisible forms of existence inside machines, such as a sound recording, a digital file, or encoded data, may all be regarded as copies if they contain expressive content, and their legal status is not different from that of traditional copies. Such copies may be 'not available directly to human beings' and may only be perceived by humans with the assistance of appropriate retrieval equipment or a suitable decoding process. [85] Copyright law does not require a copy to be directly perceived by humans in practice in every case, or to exist permanently. What it requires is the possibility of proving, retrieving, or reproducing the protected expression alleged to have been fixed in that medium by appropriate technical means.

In *Levola Hengelo,* the CJEU showed that not every form capable of being perceived by humans satisfies the requirements of copyright law. The taste of a food product can

[83] WIPO, 'Copyright' (1982) Copyright 18(9) 246; Ficsor (n 59) 96-100.
[84] Ficsor (n 59) 101-102.
[85] WIPO, 'Draft model provisions for legislation in the field of Copyright, III. Comments on the draft model provisions for legislation in the field of copyright' (1989) Document CE/MPC I/2-III 8 [39].

be directly perceived by humans, but it cannot be retrieved or reproduced as a defined and stable expression for everyone. The same product may create different perceptions among different tasters, or even for the same person at different times, making it impossible to identify precisely what the alleged expression claimed for protection is. [86] Therefore, a copy under the reproduction right must have objective identifiability, meaning the ability for the relevant subject matter to be identified, verified, and distinguished as an expression with 'sufficient precision and objectivity'.[87]

In addition, another criterion for determining the scope of the reproduction right under Article 2 of the InfoSoc Directive, developed from the CJEU's case law in Pelham I [88] is the 'recognisability of the original elements of the work in the other object'. [89] In the case of AI memorisation, this requirement will usually be satisfied, because memorisation is often detected through outputs that reproduce, exactly or almost exactly, a pre-existing work or a substantial part of that work, where the original expression of the work can be easily recognised.

However, scholars remain cautious about the possibility of applying this element in the context of copyright. [90] This is because the 'recognisable' condition was developed by the CJEU in the context of the related right of phonogram producers under Article 2(c) of the InfoSoc Directive, and not directly in the context of copyright in 'works' under Article 2(a). In that case, there was no dispute or analysis as to whether the part taken was a protected expression of a copyright work. Instead, the Court's reasoning moved directly to the scope of the related right of the phonogram producer and justified the standard of being 'recognisable to the ear' [91] by reference to the logic of protecting the producer's ability to recover its investment in the sound recording. [92] Therefore, this standard should not be transferred easily to copyright works, where the central question remains whether the part taken reflects the author's own intellectual creation.

[86] Case C-310/17, *Levola Hengelo BV v Smilde Foods BV* EU:C:2018:899 ('*Levola*')[40]-[44].
[87] *Levola* (n 86) [40].
[88] Case C-476/17, *Pelham GmbH and Others v Ralf Hütter and Florian Schneider-Esleben* ('*Pelham I*') EU:C:2019:624.
[89] as discussed in *GEMA* (n 7) [243]-[244].
[90] Bernd Justin Jütte and João Pedro Quintais, 'The Pelham Chronicles: Sampling, Copyright and Fundamental Rights' (2021) 16 JIPLP 213; Emanuilov and Margoni (n 26) 17.
[91] *Pelham I* (n 88) [37]-[39].
[92] *Pelham I* (n 88) [30], [38]-[44].

3.2 Should Memorisation Be Considered Reproduction

The previous chapters have established two foundations for the present analysis. First, memorisation in GenAI models is not merely a speculative concern, but a technically observable and provable phenomenon. Second, EU copyright law gives the reproduction right a broad and technological neutrality scope of interpretation. At the same time, through the historical development of the concept and the case law of the CJEU, a number of criteria have been developed to determine what amounts to reproduction.

This section will examine whether memorisation in GenAI models, especially LLMs, with the characteristics analysed above, satisfies the criteria for being regarded as an act of reproduction within the scope of Article 2 of the InfoSoc Directive. In other words, it examines whether, in the process of training, forming and operating AI models, there exists within the model an expression containing the author's intellectual creation, fixed in a form that can be identified with sufficient precision and objectivity.

The search for an answer to this question still appears far from settled. The debate in legal scholarship, which has moved beyond the technical issues surrounding machine learning, remains highly active. No view has clearly prevailed, and all sides are still waiting for firm and clear legal and policy decisions from competent authorities. The outcome of this debate will lead to significant changes in the liability of AI companies and in the ability of rightholders to control the use of their works.

3.2.1 The Ongoing Debate

The literature on AI memorisation and copyright remains unsettled. For the purposes of this thesis, the debate can be organised around three main approaches. The first treats model parameters as statistical correlations rather than copies, and therefore resists treating model memorisation as reproduction. The second argues that where copyrighted expression can be reconstructed from the model, memorisation may become legally relevant to the reproduction right. A third approach, which partly overlaps with the second, is based on the functional equivalence doctrine to ask whether

the ingestion or embedding of works into model parameters may perform a copy-like legal or economic function, even if the model does not contain a conventional digital file. These approaches should not be understood as rigid categories. but rather serve only as an analytical reference point for applying the technical features of memorisation to the EU law concept of reproduction.

Cooper and Grimmelmann combine technical expertise in machine learning with legal analysis of copyright to develop a set of concepts that helps 'translate' the technical phenomenon of AI model memorisation into a usable way for legal audiences. [93] Their work has been widely referred to and cited by scholars working on related issues. [94] This thesis also adopts their conceptual framework in order to use the relevant terms with clear and consistent meanings in the analysis below.

The most important aspect of their proposal lies not only in the meaning of each term, but also in the clear separation between different layers of the phenomenon, so that not every manifestation related to training data is grouped together under the label of 'memorisation'.

According to them, 'memorisation' should be used only to describe a model-level phenomenon: a situation in which a substantial portion of a specific training example can be reconstructed, almost verbatim, from the model's parameters.[95] Meanwhile, 'extraction' and 'regurgitation' are use-level and output-level phenomena. Extraction refers to the user's active attempt to obtain content that has been memorised, while regurgitation refers to the model generating a near-verbatim copy, whether or not the user intended this result. [96] This distinction is not meant to exclude prompt-triggered cases from the scope of memorisation. Rather, it emphasises that, in some cases, the prompt may determine how memorised content is revealed at the output generation stage, but it does not determine whether the model has memorised the content in the first place. In other words, 'memorisation' refers to an internal feature of the model

[93] Cooper and Grimmelmann (n 17) 142-143.
[94] EUIPO, *The Development of Generative Artificial Intelligence from a Copyright Perspective* (EUIPO 2025) 157 DOI: 10.2814/3893780.
[95] EUIPO (n 94) 158-159.
[96] EUIPO (n 94) 159-160.

formed during training, while ‘extraction’ and ‘regurgitation’ are only external manifestations of that property. [97]

3.2.1.1 Arguments That AI Memorisation Does Not Qualify as Reproduction

Many ‘no-copy’ arguments tend to attribute memorisation issues to extraction. According to this view, much of the evidence relied on to prove memorisation is, in fact, produced by giving the model a very long prefix, one that is very close to the original work, or by repeatedly providing identifying signals of a specific work. In such cases, the user is actively conditioning the model’s probabilistic prediction process, pushing it to follow the token sequence that the model has learned to treat as the most probable continuation. [98] Therefore, when the model completes the remaining part of the passage under these conditions, this counter-argument suggests that it does not prove that the model is ‘opening’ a stored copy. It only shows that the model has been guided to perform its probabilistic mechanism. On this view, regurgitation occurs only at the point of inference, rather than directly proving that the model has retained the work internally as a stable copy.

In *The New York Times v Microsoft and OpenAI*, OpenAI argued that the examples provided by The New York Times were the result of efforts to ‘hack OpenAI’s products’ and required ‘tens of thousands of attempts’ to produce abnormal results. [99] In Getty Images v Stability AI, the court also considered whether the prompt given to the AI was deliberately artificial to such an extent that it would be unlikely to arise in ordinary use.[100] Although the court did not examine this issue for the purpose of proving memorisation, it shows that the deliberate design of a prompt to trigger a particular form of output may directly affect the evidential value of that output.

---

[97] EUIPO (n 94) 168.

[98] Aline Larroyed. 'The Fallacy of the File: How the Memorisation Metaphor Misguides Copyright Law and Stifles AI Innovation' (SSRN 2025) Section 3 <http://dx.doi.org/10.2139/ssrn.5782882>.

[99] OpenAI, ‘OpenAI and Journalism’ (*OpenAI*, 8 January 2024) <https://openai.com/index/openai-and-journalism/> accessed 24 May 2026; Tyler Cowen, ‘Toothpick Producers Violate NYT Copyright’ (*Marginal Revolution*, 30 December 2023) <https://marginalrevolution.com/marginalrevolution/2023/12/toothpick-producers-violate-nyt-copyright.html> accessed 24 May 2026.

[100] *Getty* (n 6) [32]-[62].

These assessments imply that memorisation and copying occur only at the moment of output generation, rather than as a phenomenon inside the model. This leads to the conclusion that outputs may infringe where they are verbatim or substantially similar to a copyrighted work, but not the model itself. [101] However, the argument that a work-like output appears only after an unusual prompt is not sufficient to completely rule out the possibility of memorisation. It may affect the assessment of whether the use is ordinary, or the persuasive value of the evidence in a particular case, but it does not in itself exclude the possibility that protected expression had already been retained in the model.

From a technical perspective, as explained in Section 2.2, methods for proving memorisation often use repeated prompting, prefix prompting, or targeted elicitation strategies, but these are not the only methods. In particular, an individual study by GitHub's Albert Ziegler, which tested Copilot's ability to recite code, showed that Copilot 'quotes when it lacks specific context' and 'mostly quotes in generic contexts'.[102] When a file is still at a very general stage and lacks specific signals indicating what the user intends to write, the model is more likely to suggest passages that it has seen many times in the training data.[103] In this case, the absence of specific context itself causes the model to fall back on what is familiar from the training data, including non-functional code elements such as comments, formatting, naming choices, or passages associated with a particular file. The results of this study are commented that it shows a capacity for storage somewhere inside the models themselves.[104]

A stronger counter-argument focuses on the nature of model parameters and points to the ontological difference between them and the understanding of a copy that has been established over the past decades. GenAI models do not store training works in the same way as a hard drive, database, archive, or ordinary digital file. Their internal structure consists of parameters, weights, and vector representations that are adjusted during training in order to optimise predictive ability. On this basis, opposing views argue that

[101] Cooper and Grimmelmann (n 17) 145-146.
[102] Emanuilov and Margoni (n 26) 24.
[103] Ziegler (n 24).
[104] Emanuilov and Margoni (n 26) 24.

parameters, such as tokens, vectors, weights, or other elements inside the model, cannot be treated as copies of copyrighted works.

At the stage of tokenisation and embedding, protected works are transformed into token-level statistical relations and numerical representations, and it is difficult to regard them as a fixation containing expressive content. One article has compared tokens with object code to clarify this point. [105] Indeed, token IDs and embeddings may be fixed as numerical data, but token IDs mainly operate as technical labels for units of language, while embeddings - although they may reflect semantic or contextual relations learned in vector space - do not in themselves preserve the protected work as a defined expressive object. [106]

Legal debates therefore mainly focus on model weights or parameters, that is, the parts adjusted during training and said to be where patterns or memorised training data may be encoded. This is also consistent with technical studies on memorisation, which mainly analyse the possibility that training data is retained in the trained model through parameters or weights.[107] However, this approach also easily faces the objection that what exists in the weights is only distributed statistical correlations, not a work fixed in a medium. [108] Under this view, model parameters merely reflect abstract statistical correlations drawn from a large corpus, and the changes made to parameters during training can only be regarded as being influenced by, learning from, or, as some writings put it, containing 'ideas', [109] but not as storing any copyrighted expression from a specific work. They therefore cannot be treated as internal copying within the model, because copyright law normally does not treat influence, learning or abstraction as copying.

The identifiability of the copy also becomes a central point of debate, where the relationship between numerical representations and the protected work is examined.

[105] Emanuilov and Margoni (n 26) 27-28.
[106] Emanuilov and Margoni (n 26) 27-28.
[107] Milad Nasr and others, 'Scalable Extraction of Training Data from (Production) Language Models' (2023) arXiv:2311.17035, 15 < https://doi.org/10.48550/arXiv.2311.17035>.
[108] Van Lindberg, 'Building and Using Generative Models under US Copyright Law' (2023) 18 Rutgers Bus L Rev 1.
[109] Rita Matulionyte, 'Reconceptualising the Reproduction Right in the Age of AI' (2025) 56 IIC 1914, 1927, 1932 DOI: 10.1007/s40319-025-01653-x.

Digital copies, such as text files, image files, audio files, or object code, although intangible and not directly perceptible, still maintain a defined relationship with the protected subject matter. Therefore, when appropriate software or equipment is used, the content of the work in those copies can still be perceived, retrieved or reproduced as the same object in a consistent and identifiable way. By contrast, model parameters face difficulties in proving this quality, [110] as the information extracted during training is fragmented and distributed across a high-dimensional parameter space, to the point that no single parameter directly corresponds to the protected expression. Indeed, according to several scholars, the potential reproduction of training data in a trained AI model does not satisfy the requirement of identifiability and therefore cannot be regarded as a protected copy under the reproduction right. [111]

In the context of EU copyright law, where the reproduction right is broad but still linked to the notion of protected expression, the question of whether weights or parameters can be treated as a form of expression, or as a form of fixation containing objectively identifiable content, carries significant weight. It must be acknowledged that, at least with the current state of technology, the content retained in a model hardly exists in a form that can be identified or retrieved in a stable way. Although many technical methods have clearly shown that memorisation is an internal mechanism of models, reconstructing a specific work from a model still depends heavily on repeated prompting attempts, sampling conditions, and other probabilistic factors, rather than through a determinate retrieval mechanism as familiar digital copies.

For this reason, those following this line of argument consider that memorisation, at least in ordinary cases, should be understood as cognitive assimilation or information extraction rather than copying. [112] Treating memorisation as reproduction would stretch the concept of copyright beyond the scope it should have, especially where specific works cannot be identified in the parameters. [113] Some stricter views, such as that of

[110] Alex Brando, *Technological Aspects of Generative AI in the Context of Copyright, Attribution and Novelty in Generative AI Hypersurfaces* (Requested by the JURI Committee, Policy Department for Justice, Civil Liberties and Institutional Affairs Directorate-General for Citizens' Rights, Justice and Institutional Affairs PE 776.529, July 2025) 14-15.
[111] Matthias Leistner and Lucie Antoine, ‘TDM and AI Training in the European Union - From “LAION” to Possible Ways Ahead?’ (2025) 11 GRUR International 1032–1033.
[112] Zbigniew Okoń, 'Memorization of Copyrighted Works by AI Models Under EU Law' (SSRN, 2026) DOI:10.2139/ssrn.6041395.
[113] Okoń (n 112).

Aline Larroyed, criticise the equation of the model with a 'filing cabinet' containing works as a misleading metaphor that diverts legal discourse away from technical reality.[114]

3.2.1.2 Arguments In Favour of Treating AI Memorisation as Reproduction

Arguments denying memorisation as a form of reproduction are not without basis, especially where they emphasise that the content inside the model does not constitute a fixation containing the expressive content of the work, and cannot be retrieved in a stable way like other digital copies with which we have become familiar over the past four decades. [115] However, scholars and commentators who support the view that AI models contain copies seek to broaden the understanding of reproduction by treating 'AI as the new method for storing and communicating data'. [116] This broader understanding is supported by, as analysed in Section 3.2.1 above, the broad and flexible interpretation of reproduction in EU copyright law in response to technological functions, and by the requirement of fixation in the digital environment, which does not strictly require a copy to exist in a static or immutable form.

Proponents do not seek to respond directly to every element of fixation or material embodiment in the traditional sense. They argue that it is not necessary to identify a separate physical or spatial location where the work is stored. They acknowledge that model parameters do not store information through a 'one-to-one encoding' or 'byte-to-byte' mechanism like familiar digital files, and that memorised content may be distributed and overlapped across a very large number of parameters. At present, there is also no tool that can identify exactly which parameter has encoded which part of the work. [117] However, fixation and copies in copyright law should be understood as functional rather than formal, [118] and under this approach, the requirement of fixation may still be satisfied.

[114] Larroyed (n 98).
[115] Matulionyte (n 109) 1935.
[116] See Matthew Lindberg, 'Applying Current Copyright Law to Artificial Intelligence Image Generators in the Context of Anderson v. Stability AI, Ltd' (2024) 15 Cybaris 37, 43; Okoń (n 112) 4-5.
[117] Cooper and Grimmelmann (n 17) 191-194.
[118] Cooper and Grimmelmann (n 17) 184.

Parameters, weights and other internal representations inside AI models are, in substance, a form of material existence. Similar to other machine-readable configurations in the digital environment, they are a concept of an existence shaped within human understanding. A comparison with ZIP compression has been used to illustrate this similarity. A compressed file also does not contain the original data byte-for-byte, and the data in its stored state cannot be directly understood by humans. However, because the original data can be reconstructed through the corresponding algorithm, the compressed file can still be regarded as a meaningful embodiment for the reproduction right. [119] Of course, the two situations should not be fully equated, because model parameters do not operate like a ZIP file with a determinate and stable decompression algorithm. However, this comparison shows that copyright law should not attach fixation to the ability to precisely locate each expressive unit corresponding to each storage unit, especially in the context of new technologies such as AI. If it can be shown that a substantial part of the protected expression in a specific work has been encoded in the model and can be reconstructed with a near-exact level of accuracy, then the absence of a clear 'location' [120] in the parameters is not necessarily a decisive reason to deny the existence of material embodiment somewhere inside the model.

This point further raises the questions of the meaning and purpose of the requirements of identifiability, traceability, or retrievability when determining the existence of a copy. Are these criteria meant to require that a copy must, in all circumstances, be perceived or retrieved by everyone? Or are they meant to establish a real and objective fact that an act of storing part of a copyrighted work has occurred?

The first understanding is probably too narrow. A copy does not lose its status as a copy merely because an ordinary user does not have the necessary equipment, software, technical knowledge, or decryption key to perceive the content contained in it. A music disc, an encrypted file, a record on a storage device, or an ancient text that can be read

[119] Cheng Lim Saw and Bryan Zhi Yang Tan, 'Unpacking Copyright Infringement Issues in the GenAI Development Lifecycle and a Peek into the Future' (2025) 58 CLSR 106163, 6 DOI: 10.1016/j.clsr.2025.106163.
[120] Cooper and Grimmelmann (n 17) 185.

only by certain people and under certain conditions may still be a carrier of copyrighted expression. [121]

Under the second understanding, retrievability, traceability, or identifiability does not necessarily mean the ability to recover the work in a deterministic way. Nor are these requirements intended to ensure that everyone can access or enjoy the work in every circumstance. This understanding does not undermine the requirement of objective identifiability. It only clarifies the role of that requirement in serving proof and legal attribution.

Indeed, retrievability also does not mean, in absolute terms, that the work must be reproduced successfully in every attempt. Even systems regarded as deterministic in practice may fail under unusual technical conditions. [122] Similarly, in the AI context, the fact that a model does not generate the same output in every prompting attempt cannot by itself mean that it lacks determinacy. What matters is whether the work can be reconstructed with a sufficiently accepted and reliable probability, so that the result cannot reasonably be attributed to mere coincidence. The feasibility of retrieving the expressive elements of a work with a sufficiently high probability for evidential purposes has been clearly shown through the machine learning studies discussed above. [123]

Therefore, although the most effective current method for observing memorised training data is prompting and observing the output, this does not change the nature of the phenomenon or remove the evidential value of this method. We cannot directly see training data inside the model's parameters, but it can be confirmed through indirect

[121] '[...] it does not matter whether we understand the internal code or not. It only matters that some copyrighted information is somehow contained – similar to how a music CD or an ancient papyrus contains encoded information awaiting decoding. Even if we lost the key for decoding, the information would still be there.' Sebastian Stober and Tim W. Dornis, 'Generative AI Training and Copyright Law' (2025) arXiv:2502.15858 <https://doi.org/10.48550/arXiv.2502.15858>.

[122] 'even "deterministic" processes have a little non-determinism baked into them. A jukebox in factory condition has a small but non-zero probability of malfunctioning: maybe there will be a power surge at exactly the wrong moment, and the tonearm will never make contact with the record. The algorithm for converting an MP3 file to an audio signal is deterministic— but there is always a non-zero (if tiny) probability that cosmic rays will corrupt the computer's memory in a way that overwrites the audio data with noise.', Cooper and Grimmelmann (n 17) 198.

[123] As provided in Section 2.2.

evidence - just as many physical phenomena can only be known through measurement or intermediate signs, without this giving us a reason to deny that they exist. [124]

Therefore, the instability in making protected expression appear at the output stage should be treated as an obstacle that complicates proof, rather than as a factor excluding the recognition of reproduction at the model level. In the context of AI models, identifiability may be understood as the ability to prove, sufficiently and convincingly, that a substantial part of a copyrighted work has been encoded in the model and can be made to appear through an appropriate technical process. If prompting or other available measurement methods produce an output containing the whole or a substantial part of a protected work, or show that the output is substantially similar to a pre-existing original work, this clearly indicates that the work systematically reproduced by AI models can be retrieved. [125]

Cooper and Grimmelmann further emphasise another point: even if AI models are said to learn only 'patterns' or 'statistical correlations', this is not enough to exclude memorisation from the scope of reproduction. Concepts such as 'pattern' or 'statistical correlation' are broad and abstract. A pattern may be only a general rule drawn from many training examples, in which case the model mainly generalises. At the same time, however, a pattern may also be so specific that it reflects a particular training example, such as the sequence of words in a passage, the layout of an image, or distinctive expressive choices that cannot be explained simply as a general rule. [126] This can be compared to the boundary between 'idea' and 'expression' when determining which elements of a work are protected. A plot at a general level may be only an idea, but if that plot is sufficiently specific, complex, and marked by creative choices, it may become a protectable expression. [127] Similarly, if GenAI models can still reproduce a

[124] 'True, we cannot observe the memorized training data directly in the model's parameters—but neither can we directly observe black holes, ultraviolet light, or electric fields. We can confirm their existence through indirect measurements—detecting certain types of nearby radiation, using specialized sensors, and observing behavior of charged particles, respectively. In the same way, extraction of memorized training data is a kind of indirect measurement.' Cooper and Grimmelmann 'The Files are in the Computer - On Copyright, Memorization, and Generative AI' p 169

[125] Saw and TAN (n 119) 6.

[126] Cooper and Grimmelmann (n 17) 186-187.

[127] Judge Learned Hand's point was that, as more concrete incidents of a play's plot are left out, the resulting pattern becomes increasingly general. At some point, it falls outside copyright protection, but before that level of abstraction is reached, there may still be sufficiently concrete layers of the plot that remain within the realm of protectable expression. See *Nichols v Universal Pictures Corp 45 F 2d*

substantial portion of training data, calling the learned content ‘patterns’ does not change the fact that protected expression has been retained at a specific level.

The broader view in favour of treating AI memorisation as reproduction is also reflected in some recent practical developments. In *GEMA v OpenAI,* the Munich Regional Court considered that the model’s ability to output fragments of lyrics was an indication that those fragments had been stored in the model, and that memorisation could therefore constitute reproduction.

3.2.3 Reconceptualizing the ‘Reproduction’ Definition

Dr Rita Matulionyte, a scholar working on copyright and AI, also follows a line of argument that supports the interpretation that the reproduction right covers AI model memorisation, and that excluding memorisation entirely from the reproduction right may weaken the effectiveness of this right in the AI environment. However, Matulionyte does not directly address whether each element of memorisation satisfies the established criteria for reproduction. Instead, she focuses on whether the ‘ingestion’ or ‘embedding’ of protected content into model parameters performs the same legal and economic function that the reproduction right is intended to address. [128]

This approach relies on the doctrine of functional equivalence, a doctrine frequently applied in law [129]. Although this doctrine does not yet have a clear meaning in the context of lawmaking, it is commonly understood as requiring that ‘the law should treat a similar activity in a similar manner’, even where the form or technology through which the activity is carried out differs. [130] This doctrine is ‘closely connected to the principle of technological neutrality’, a principle already present in copyright law.

119 (2d Cir 1930); Judgment in *Laras Tochter* shows that a plot may initially appear to be an unprotectable idea or a general narrative pattern, however, where the plot is developed through sufficiently specific characters, relationships, settings and key incidents, it may no longer remain a mere idea, but become a protectable expression. *Laras Tochter*, Case No I ZR 65/96 (Bundesgerichtshof, 29 April 1999), BGHZ 141, 267.

[128] Matulionyte (n 109).

[129] See also Chris Reed, ‘Online and Offline Equivalence: Aspiration and Achievement' (2010) 18 IJLIT 248 <https://doi.org/10.1093/ijlit/eaq006>.

[130] Johan David Michels and Christopher Millard, 'The New Things: Property Rights in Digital Files?' (2022) 81(2) The Cambridge Law Journal 323 <https://doi.org/10.1017/S0008197322000228>; ‘identical services should in principle be regulated in the same way, regardless of their means of transmission.’ European Commission, ‘Principles and guidelines for the Community's audiovisual policy in the digital age’ (Communication) COM (1999) 657 final, fn 17.

When digital technology emerged, EU law did not require digital copies to be perceptually identical to physical copies. Instead, it recognised that digital storage may still constitute reproduction because it performs the same function of storing and exploiting content.

Under this understanding, even if AI models do not contain 'digital copies as we have understood them', they may still contain 'the representation of works' that is '"functionally equivalent" to copies of work'. Therefore, the reproduction right should be extended to cover this new form of use where it corresponds to 'the core function of the reproduction right and copyright law more generally'. [131]

Therefore, the 'ingestion' of the content of a work into an AI model, whereby that content is encoded in parameters and may serve further reuse purposes, can be regarded as functionally equivalent to storing a copy. This has a basis in the 'core function' that the law attaches to reproduction. Reproduction is seen as a fixation that allows a stored digital copy to be treated as reproduction, where it 'permits indirect communication to the public' or 'further reproduction (the making of a new fixation)', that is, the reuse of the work.[132]

The proposal to apply the doctrine of functional equivalence broadens the analysis to the practical effects of technologies, such as their legal and economic functional similarity, [133] which helps reduce the difficulty of proof, because it does not require locating a copy according to a fixed set of criteria. Output is not excluded from the analysis; it may still be important evidence, but it is not necessarily the only route of proof. In theory, if a technical method can show, now or in the future, that part of a copyrighted work is retained in the model in a way that can be reproduced or exploited, functional equivalence allows the function of that representation to be assessed directly, without being tied to a formal checklist. What matters is to examine, in each specific case, what the representation inside the model does, and whether it retains protected

[131] Matulionyte (n 109) 1935.
[132] Ficsor (n 59) 94-95, 444-445, reflecting a position later confirmed in the Agreed Statements to the WCT 1996, in the interpretative practice of the Berne Convention, and reinforced by the CJEU case law on RAM copies.
[133] Matulionyte (n 109) 1936.

expression in a way that may serve the reuse, retrieval, or exploitation of the content of the work.

On the other hand, this doctrine should not be understood as treating every act of feeding data into machine learning as reproduction. The issue of reproduction should arise only where the subject matter retained and exploited is protected expression, not every regularity in the training data, such as abstractions, factual correlations, general stylistic patterns, or unprotected ideas.

Beyond the doctrinal debates analysed above, many scholars have also expressed concerns about the practical effects that an official conclusion on this issue may have on stakeholders in the real market, especially in relation to the application of exceptions, in particular Article 4 of the DSM Directive. Notably, even scholars who accept, to some extent, that memorisation may theoretically be regarded as an act of reproduction also express practical concerns that an overly strong extension and protection of the reproduction right may risk disrupting the copyright balance that the EU copyright acquis seeks to establish. [134] For this reason, a more careful and refined mechanism than the existing one may be needed. The specific issues concerning exceptions and policy will be discussed further in Sections 5 and 6 below.

Finally, it should be emphasised once again that the reproduction right in copyright law has never been immutable. The history of the reproduction right is a history of adaptation to new technologies. When digital technologies emerged, this right was repeatedly extended, through legislative amendments and the case law of the CJEU, to cover new objects and phenomena, from sound recordings, copies in computer memory, works embodied in computer programs, representations of works in binary form, and cached copies, to digital fragments. [135] Today, AI raises a similar challenge, as works are encoded in models in a way that has not existed before. This does not mean that they should be excluded from the scope of reproduction. Rather, the definition of

[134] See, eg Péter Mezei, 'Memorization and Generative AI - A Persistent Issue with Copyright Consequences?' (SSRN 2025) <http://dx.doi.org/10.2139/ssrn.5404367>, Emanuilov and Margoni (n 26), Okoń (n 112).

[135] This issue had already been addressed in the debates between 1982 and 1996, during the period leading up to the Diplomatic Conference, and has been repeatedly reaffirmed throughout subsequent stages of technological development.

reproduction itself may need to adapt to the new context. There is no reason not to expect this right to be interpreted flexibly once again, in order to ensure the high level of protection for authors' copyright that EU copyright law seeks to achieve. [136]

3.3 GEMA v OpenAI: A Notable Decision from the Munich Regional Court

3.3.1 Case Facts

The decision in *GEMA v OpenAI*[137] may be regarded as one of the first rulings within the EU directly addressing the issue of memorisation in AI models. It attracted significant attention from legal scholars, particularly in light of the earlier decision in Getty Images v Stability AI in the UK.[138]

The claimant, GEMA, a German collective management organisation, brought proceedings against OpenAI entities alleging copyright infringement in relation to nine song lyrics. GEMA's claim operated on two levels: it argued that the lyrics had been memorised and thereby reproduced within the language models themselves, and that the chatbot's outputs constituted additional acts of reproduction and making available to the public. GEMA demonstrated that, upon simple prompts, the chatbot reproduced verbatim or substantial recognisable parts of the lyrics without authorisation. It was undisputed that the disputed works had been included in the training data.

The Munich Regional Court held that[139] the memorisation of works within AI models constitutes reproduction under Section 16 UrhG and Article 2 of the InfoSoc Directive. It found OpenAI liable for infringement of reproduction and communication rights, issued an injunction against further reproduction of the works in Germany.

---

[136] InfoSoc recital 9.
[137] See the final judgment, LG Munich I, Endurteil v. 11.11.2025 – 42 O 14139/24 'Unzulässige Vervielfältigung durch Memorisierung von Werken im und durch KI-Sprachmodell (Unauthorized reproduction through memorization of works in and through AI language models)' (*Bayern Recht*, 2025).<https://www.gesetze-bayern.de/Content/Document/Y-300-Z-GRURRS-B-2025-N-30204?hl=true> accessed 24 May 2026.
[138] See Mrs Justice Joanna Smith's judgment in *Getty* (n 6) [599]-[601].
[139] *GEMA* (n 7).

3.3.2 Judicial Reasoning and Academic Reactions

To analyse the legal issues in the case concerning the training and use of LLMs, which is OpenAI's model in this case, the Munich court divided the lifecycle of the model into three stages: pre-training, training, and output generation.

> 'A distinction must be made between the following successive phases:
> (1) the extraction and conversion of the training material into a machine-readable format, the creation of the training data material,
> (2) the analysis of the data material and its enrichment with meta-information, the training of the model, and
> (3) the subsequent use of the trained model through prompts and outputs.' [140]

The operation of GenAI may involve successive acts of reproduction and, therefore, several stages may raise copyright issues. The making of copies at the stage of collecting and preparing training data is often discussed as potentially falling within the TDM exception under Article 4 of the DSM Directive, and it was not the issue brought before the court by the claimants. The case instead focused on copies generated at the output stage, which also served as evidence for looking back and questioning what 'remained inside' the model during and after the training stage.

Although this division of stages has been criticised for not accurately reflecting the technical nature of machine learning, since it groups the steps involved in 'training' into a single phase and does not fully capture the gradual nature of model learning or the possibility that some early memorisation may disappear as generalisation develops, [141] it remains useful from a copyright law perspective. It separates different groups of acts that may relate to the making and use of copies of training data, and it also assists in the assessment of whether exceptions concerning copies made in the process of text and data mining may apply.

---

[140] English translation provided in 'Copyright Infringement in Form of a Reproduction of Preexisting Works in a Large Language Model' (2026) 75(4) GRUR International 371 <https://doi.org/10.1093/grurint/ikag009>.

[141] A. Power and others 'Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets' (arXiv 2022) <http://arxiv.org/abs/2201.02177> accessed 24 May 2026.

Considering the evidential issue and the analysis of technical elements, in paragraphs 169 to 171, the court relied on computer science research showing that training data can be memorised in the model, thereby recognising the objective existence of this phenomenon. In this specific case, memorisation could be proved by comparing known training data with outputs generated from sufficiently long 'simple prompts', and by assessing evidence that the overlap in lyric lines could not be explained by chance, given the length and complexity of the lyrics.

The fact that output containing protected subject matter could be elicited by simple prompts significantly strengthened the evidence. It suggested that this was not merely a rare example, but something that could appear with a sufficiently high probability in the ordinary use of an average user [142]. This reasoning is quite consistent with the suggestion made by some scholars concerning 'reasonable efforts' in proving memorisation. [143] However, unlike the English court in Getty Images v Stability AI, the Munich court focused on the persuasiveness and legitimacy of the evidence submitted by the claimants, without clearly stating whether other efforts to steer AI models into producing copyrighted content would completely remove the evidential value of such outputs.

In paragraph 180 of the Judgment, the court provided a provisional definition of reproduction: '[...] any physical fixation of a work that is capable of making the work perceptible to the human senses in any way, either directly or indirectly, is considered to be a reproduction.' On that basis, the court examined whether the phenomenon of AI memorisation in the case could satisfy the criteria for being treated as a copy.

First, the Court affirmed that there was a fixation containing protected content. The fact that the work appeared in the form of probability values or distributed parameters was not relevant to whether it could be regarded as a fixation. The court compared this form to a lossy MP3 file or a progressive JPEG file, where information is incomplete or distributed within the file but can still be restored and recognised. This comparison is

[142] *GEMA* (n 7) [172]-[175].
[143] Linda Kuschel and Darius Rostam, 'If It Looks Like a Duck On the Munich Regional Court's ruling in GEMA v. Open AI' (*Verfassungsblog*, 19 November 2025) <https://verfassungsblog.de/munich-regional-courts-ruling-in-gema-v-open-ai/> accessed 24 May 2026.

similar to the ZIP file example discussed earlier in this thesis and supported by scholars who regard model parameters as a copy. Such comparisons have also been criticised as an 'oversimplification', because they do not pay sufficient attention to the difference in nature and purpose between these objects, thereby shifting the focus away from the proper analysis of the reproduction right. [144]

As to retrievability through prompts and indirect perceptibility through the display of output, the court considered that these elements fulfilled the requirement of identifiability. In addition, the court did not limit the evidence to actual outputs, but also recognised the possible relevance of other types of evidence, such as tests based on entropy and perplexity parameters to assess the degree of certainty with which a model reproduces an output. [145] The court emphasised that 'For the purposes of reproduction under copyright law, there is no need to determine how memorisation works in detail [...] The decisive factor is that the song lyrics that served as training data are reproducibly contained in the model and thus embodied in it.' This shows that the Munich court interpreted reproduction under the InfoSoc Directive and the UrhG broadly, in line with the wording 'by any means and in any form' [146] and the principle of technological neutrality, thereby recognising AI model memorisation of training data as reproduction at the model level.

The Munich court pushed this reasoning further by clearly separating the ability to make a work perceptible from the public's ability to access the work when assessing the elements of reproduction. According to the court, the fact that a work can only be made perceptible in a limited way does not exclude the existence of reproduction. It only means that the work has not been made public, while public accessibility is not a condition of the reproduction right. In other words, the reproduction right does not require the copy to be directly or easily accessible to the ordinary public. It is sufficient that there is a possibility, by some means, for humans to perceive it, directly or indirectly, so as to confirm the existence of that copy. This approach had already been

[144] 'The crux of the problem of assessing memorization under copyright law does not appear to lie in whether the scope of the reproduction right covers the making of a copy capable of "indirect" perception, but rather in whether reproduction in the normative sense encompasses the making of copies not intended for perception, within a process not designed to record works for subsequent playback, and dependent to a significant extent on probabilistic effects.' Okoń (n 112) 6-7.
[145] *GEMA* (n 7) [171].
[146] InfoSoc art 2.

mentioned by Cooper and Grimmelmann in their work and has been considered 'correct insofar as the question of whether a reproduction exists is normative and not merely technical', [147] confirming that the important point lies in the extractability of specific works.

GEMA v OpenAI shows the need to distinguish clearly between two stages of copyright infringement in AI memorisation. First, from the moment a work is 'engraved' inside the model, this may already be regarded as the making of a copy. Second, the display of protected expressive elements in the output to users should be assessed separately as another act of reproduction and, at the same time, as an act of communication to the public. It should also be noted that the court did not conclude that every model contains or copies its entire training dataset in all cases. Although it recognised AI memorisation as an inherent feature of current GenAI, a claimant still needs to prove the existence of a copy in each individual case in order to establish a copyright infringement claim.

Although this is only a first-instance judgment and has been appealed, and although there is still no conclusion at the EU level, it is the first decision in Europe to formally recognise AI memorisation as an act relevant to copyright. The reasoning of the Munich court has been welcomed by scholars who consider memorisation to fall within the scope of the reproduction right from the stage at which information is stored in the model, and who argue that rightholders' rights should be strongly protected in the current AI context. For instance, Silke von Lewinski praised the decision in her article, stating that 'it is remarkable in its depth, clarity, rigorous analysis, and pertinent observations, and should be followed as a model by other courts including at higher instances.'. [148] KODA, a music rights organisation in Denmark, also expressed strong support, stating that the judgment provided both moral encouragement and support for KODA's own arguments in its case against another AI company. [149] [150]

[147] Arne Radeisen, 'Open Foundation Models and TDM Exceptions to Copyright – Building Blocks for an AI Ecosystem'(2026) 75 GRUR International 224 <https://doi.org/10.1093/grurint/ikag002>.
[148] Silke von Lewinski, 'GEMA v. OpenAI: The Munich Regional Court Presents an Exemplary Judgment on a Well Conceived Test Case' (2026) 75(4) GRUR Int 340.
[149] KODA, 'Historic ruling: OpenAI found guilty of exploiting music' (*KODA*, 11 of November 2025) <https://koda.dk/en/about-koda/news/historic-ruling-openai-found-guilty-of-exploiting-music > accessed 24 May 2026.
[150] At the end of 2025, KODA filed a lawsuit against Suno, a US AI company, concerning similar allegations of copyright infringement.

4 Applicable Exceptions and Limitations

In order to ensure a balance between the interests of authors, rightholders and the public interest, EU copyright law not only grants exclusive rights to rightholders, but also establishes a system of exceptions and limitations across different legal instruments. This system is constructed as an exhaustive list, allowing certain acts of reproduction to be carried out without the author's authorisation and without giving rise to legal liability. These exceptions are based on situations in which the use of a work can be justified by legitimate reasons recognised by society, or by policy objectives pursued by legislators, here the EU and the Member States. [151]

After the analysis in the previous chapters, it can be assumed that, where AI memorisation has been proved to be a real phenomenon and recognised as an act of reproduction, the next question is whether the existing exceptions provided by EU copyright law are sufficient to justify the retention or reproduction of protected works in and through commercial GenAI models.

Within the scope of this thesis, this section will focus only on the exceptions most directly relevant to the current issue, namely the exception for temporary copies under Article 5(1) of the InfoSoc Directive, the exception for incidental inclusion under Article 5(3)(i) of the InfoSoc Directive, and the text and data mining exception for general purposes under Article 4 of the DSM Directive.

4.1 Exceptions under Article 5 InfoSoc Directive

4.1.1 Temporary Acts of Reproduction under Article 5(1) InfoSoc

Article 5(1) of the InfoSoc Directive is also often invoked in debates on AI, because it is a mandatory exception intended to exempt technical copies that exist within a technological process. The exception for 'temporary acts of reproduction' is understood as aiming to 'exclude temporary acts of reproduction "which technology dictates"'. [152]

[151] See Lucie Guibault, *Copyright Limitations and Contracts. An Analysis of the Contractual Overridability of Limitations on Copyright* (Dalhousie University Schulich School of Law 2002) 27-110; Tito Rendas, *Exceptions in EU Copyright Law: In Search of a Balance Between Flexibility and Legal Certainty*, (Kluwer Law International BV 2021) 2-10.
[152] AG Trstenjak (n 58) point 94.

In order for this exception to apply, all the conditions set out in Article 5(1) must be satisfied cumulatively, and the exception must be interpreted strictly. In particular:

> *'- the act is temporary;*
> *- it is transient or incidental;*
> *- it is an integral and essential part of a technological process;*
> *- its sole purpose is to enable a transmission in a network between third parties by an intermediary or a lawful use of a work or protected subject-matter; and*
> *- the act has no independent economic significance*.' [153]

AI training or output generation involves many steps, from the preparation of pre-training data and the input of data into the system, to repeated stages of training, each of which requires technical copies to be made. Some of these steps create small copies, such as temporarily loading data into RAM or GPU memory, and these may be regarded as temporary and essential for computation. Article 5(1) exempts such copies only where their sole purpose is to support transmission through an intermediary in a network or to enable a lawful use of the work. However, if GenAI memorisation of training data is treated as an inherent feature of the model, then the purpose of making copies in the AI training process is to store and exploit content from those copies themselves.

In addition, according to the CJEU, for an act of reproduction to be considered temporary, the copies are expected to be automatically deleted, by an automated process, after they have fulfilled their purpose of enabling the completion of the technological process. [154] Memorisation may later lead to observable outputs, often in answers displayed on audiovisual devices. This means that the copy does not only remain inside the system, but also results in the making of a copy containing the protected expression of the work fixed on a medium at the end of the technological process. Under the case law clearly established in Infopaq I, that final act of reproduction cannot be regarded as a temporary act of copying.[155]

[153] See Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening* EU:C:2012:9 [25].
[154] *Infopaq I* (n 68) [62]-[66].
[155] *Infopaq I* (n 68) [67]-[70].

In fact, this condition directly excludes AI model memorisation, whether or not a copy appears at the output stage. Once information has been written into the system, it exists alongside the existence of the AI model itself. [156] AI companies have themselves acknowledged that, even if they wish to delete specific data after it has been incorporated into the model, this is practically impossible. Because of the fragmented, overlapping, and not directly retrievable nature of parameters in neural space, current technology cannot reverse the training process for individual pieces of content. [157]

4.1.2 Incidental Inclusion under Article 5(3)(i) InfoSoc

Article 5(3)(i) of the InfoSoc Directive is left optional for Member States to decide whether to implement it in their national copyright laws. As a result, this exception is fragmented in practice across the EU. It permits reproduction in the case of 'incidental inclusion of a work or other subject-matter in other material'. This exception was invoked by OpenAI in the litigation with GEMA, arguing that even if an act of reproduction had occurred, it resulted only from the works being incidentally present in the training dataset, rather than being the object intentionally exploited by OpenAI. In academic discussions on this issue, commentators have also raised the possible application of Article 5(3)(i).

Since the CJEU has not directly interpreted this provision and the InfoSoc Directive does not provide a definition, the wording of the provision, its context of application, and the objectives pursued by EU copyright law [158] suggest that it is intended to cover situations where the making of a copy occurs unintentionally while carrying out an act directed at another object, and where such use does not affect the interests of the rightholder.

There must be a 'material' that serves as the main object and the centre of the act of use. The second work or other subject matter appears only incidentally and unintentionally becomes part of the first material, thereby, becomes a copy stored

[156] Okoń (n 112).
[157] *GEMA* (n 7) [84].
[158] The CJEU has laid down the principle for interpreting exceptions, holding that where Directive 2001/29 does not define the concept 'those concepts must be defined having regard to the wording and context of Article 2 of Directive 2001/29, where the reference to them is to be found and in the light of both the overall objectives of that directive and international law' *Infopaq I* (n 68)[31]-[32].

within that material. A simple example is where a person films or photographs a street scene, but a copyrighted painting is incidentally captured within the frame.

The element of 'incidental' plays an important role and must be interpreted strictly so that it is not invoked too broadly. Reference may be made to the test of the German Federal Court of Justice in Moebelkatalog, according to which, in order to be considered 'incidental', the work must be 'insignificant' in relation to the main object of the communication.[159] This 'insignificance' is shown where the copyrighted work is merely a subordinate element, an unintended presence, makes no contribution to the creation of the main material, and is not exploited as a part with independent value. In other words, the copyrighted work could be replaced by other subject matter or removed entirely without any effect. [160] Moreover, this element must be assessed from the perspective of an objective average observer, who would not regard the work as part of the overall concept and would be unable to identify even the slightest relationship with the main object of use. [161]

This understanding was applied by the Munich court when rejecting OpenAI's argument in the case mentioned above. The court emphasised that the decisive factor does not lie in the quantitative proportion of the copyrighted work within the main material, but in the nature of the relationship between them. Applied more generally to GenAI memorisation, where a protected work is collected as part of what contributes to the value of the training data and is then retained inside the model with the possibility of reappearing at the output stage, it is no longer an incidental subordinate element in another material. Rather, it is part of a deliberate process of data exploitation serving the operation of the AI system.

It should be noted that 'incidental inclusion' cannot be turned into a liability shield merely by arguing that the user did not intend to exploit the work. The focus remains on an objective assessment, as reflected in the test developed by the German court.

[159] Original wording: 'Nach dem Wortlaut der Schrankenbestimmung ist vielmehr weitergehend erforderlich, dass das Werk im Verhältnis zum Hauptgegenstand der Wiedergabe unwesentlich ist' *Möbelkatalog* (Furniture Catalogue), Case No I ZR 177/13 (Bundesgerichtshof, 17 November 2014) [26].
[160] *Moebelkatalog* [27].
[161] *Moebelkatalog* [28-31].

However, in controlled and professional contexts of use, the user of the material is expected to exercise a reasonable level of care in selecting and checking what is included in the main material. Therefore, just as a painting actively placed in an interior design advertisement in Moebelkatalog could not be treated as incidental, the inclusion of a work in the training pipeline of an AI provider that knows, or ought to know, of the risk of memorisation is difficult to justify as incidental. Accordingly, this exception cannot apply to the scenario of AI model memorisation, and more broadly, to AI training in general.

4.2 Text and Data Mining Exception under Article 4 DSM Directive

The text and data mining (TDM) exception in the DSM Directive plays a central role in debates on AI training in the EU. While in the United States, 'fair use' is often invoked by AI companies as a defence, then in the EU the focus is usually on the TDM exceptions, especially Article 4 of the DSM Directive. Unlike Article 3, which is limited to research organisations and cultural heritage institutions for the purpose of scientific research, Article 4 does not limit the category of beneficiaries, as long as the act is carried out 'for the purposes of text and data mining' and the rightholder has not validly reserved the rights.

Article 2(2) of the DSM Directive defines TDM as 'any automated analytical technique aimed at analysing text and data in digital form' in order to 'generate information which includes but is not limited to patterns, trends and correlations'. From a technical perspective, [162] a TDM process usually requires the making of at least one or more intermediate copies, [163] for example, during data collection, format conversion, temporary storage, or data preparation, so that the machine can analyse it. For this reason, the TDM exception was introduced to legalise the reproductions and extractions necessary for this analytical process, insofar as they serve the purpose of TDM.

[162] For process-based descriptions of TDM, see Christophe Geiger, Giancarlo Frosio and Oleksandr Bulayenko, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects* (In-Depth Analysis requested by the JURI Committee, European Parliament Policy Department for Citizens' Rights and Constitutional Affairs, PE 604.941, 2018) 5.
[163] Geiger, Frosio, and Bulayenko (n 162) 6.

The core point of this definition is that the work is used as the object of an automated analytical process, while the expected result is information, patterns, trends or correlations, not the retention or re-exploitation of the protected expression of the work itself.

The objective of the legislators in drafting Article 4 of the DSM Directive was to facilitate technological development, including AI and machine learning, as clearly recognised in European Parliament documents. [164] This is also reflected in the TDM compliance obligation laid down in Article 53 of the AI Act. Therefore, it would not be entirely accurate to say that the EU legislature did not consider AI at all when designing the TDM exception. Rather, the better view is that it mainly had in mind the types of predictive AI that were prominent at that time, involving automated analysis and machine learning using large amounts of data, but did not anticipate the explosive rise of GenAI or the complexity of its operation, especially memorisation and the ability to reproduce protected expression at the output stage.[165]

For this reason, the debate has gradually shifted to which acts in the entire technical chain of AI training can be covered by the TDM exception, and where its limits lie. An AI training process consists of several different stages. Some steps, such as data collection, corpus preparation, format conversion, or the making of technical copies for data analysis, may be closer to the scope of Article 4 of the DSM Directive. [166] By contrast, if the training process results in protected expression being retained in a stable way in the trained model and capable of being reproduced later, that act can hardly be described merely as automated analysis aimed at generating information, patterns or correlations. Reference may be made to the approach of the German court, which clearly distinguished between 'information hidden in the data' and the 'content of intellectual creation'. [167]

---

[164] Tambiama Madiega, 'Copyright in the Digital Single Market' (Briefing: EU Legislation in Progress, European Parliamentary Research Service, PE 593.564, June 2019) <https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2016)593564> accessed 24 May 2026.

[165] European Parliament resolution of 10 March 2026 on copyright and generative artificial intelligence – opportunities and challenges (2025/2058(INI)).

[166] *Robert Kneschke v. LAION e.V.*, Case No. 310 O 227/23 (Hamburg Regional Court 27 September 2024) ('*LAION'*) [194]-[204].

[167] *LAION* (n 166) [76]-[80]; *GEMA* (n 7).

Some arguments supporting the application of Article 4 of the DSM Directive to AI training usually start from the description of AI as a data analysis tool: works are fed into the system only for automated processing, extraction of patterns or correlations, and the technical copies serving that process are then removed. That description may fit classic AI models, but it becomes incomplete when applied to GenAI - an AI system that the European Parliament recognised in its 2026 report as 'unlike other AI systems'.[168]

Even before considering memorisation in the narrow sense, generative models cannot be understood entirely as ordinary TDM tools. Their generative function remains statistically and functionally dependent on the training corpus, because the model learns a high-dimensional distribution shaped by the training examples and then generates output by sampling from that learned structure. [169] Memorisation is not the only phenomenon showing that GenAI has a deep relationship with training data. Rather, it is the clearest evidence revealing that relationship and its real potential for copyright infringement. It should be emphasised that this observation does not mean that every statistical influence from training data amounts to reproduction. However, where protected expression is retained in the model in a way that can later be reproduced or exploited, it has moved outside the scope of TDM.

The legislative history of Article 4 of the DSM Directive supports this understanding. In the Commission Proposal of 2016, the original TDM exception was not aimed at commercial practices, but was primarily designed for scientific research. [170] When the European Parliament amended the draft in 2018, Article 3a, the predecessor of the current Article 4, extended the TDM exception to other purposes, but still set a clear

[168] 'GenAI is a type of AI that, unlike other AI systems designed to primarily classify or predict, generates new content, such as text, images, music, videos and code, on the basis of training using very large datasets from which they learn patterns and structure', European Parliament (n 165) point J.

[169] Relevant technical analyses can be found in policy documents and studies commissioned by the European Parliament, see, eg Brando (n 110); Nicola Lucchi, *Generative AI and Copyright: Training, Creation, Regulation* (Study requested by the JURI Committee, Policy Department for Justice, Civil Liberties and Institutional Affairs Directorate-General for Citizens' Rights, Justice and Institutional Affairs, July 2025) 37.

[170] European Commission, 'Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market' COM (2016) 593 final, Impact Assessment section.

limit: the exception 'shall not be used for purposes other than text and data mining' [171] In the final text, Article 4(1) and (2) of the DSM Directive, although changed in wording, did not remove the limitation on the purpose for which the copies made may be used.[172] It therefore cannot be presumed that the legislature intended to prioritise commercial AI activities to such an extent that the TDM exception could be stretched to cover acts that are no longer TDM, which is precisely what the legislature had clearly sought to avoid from the beginning. [173]

Some scholars and commentators support a broader reading of Article 4 of the DSM Directive, or at least argue that AI training and memorisation should not be excluded from TDM too quickly. The central argument of this group does not always deny the copyright risks of memorisation altogether. Rather, it often starts from concerns about the practical effectiveness of the TDM exception. They are concerned that, if every stage in which a model temporarily 'memorises' data is treated as falling outside TDM, the exception may risk becoming unworkable for machine learning in practice, contrary to the requirement, set by the CJEU itself, that exceptions must be interpreted in a way that ensures their 'effectiveness and purpose'. [174] They consider Article 4 as a provision that should cover commercial GenAI development in order to promote AI technology advancement. [175] There are also warnings that equating 'memorisation' with copyright reproduction may create doctrinal drift, and that the inability to rely on the TDM exception may weaken the balance of the EU copyright acquis while disadvantaging the EU's economic and technological development objectives. [176]

The views above reflect a real concern about the risk of weakening the very mechanism that the DSM Directive introduced to support automated analysis and innovation. However, arguments based on the practical effectiveness of the TDM exception must

[171] European Parliament, 'Amendments adopted by the European Parliament on 12 September 2018 on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market' P8_TA0337 (12 September 2018).
[172]Article 4(1) DSMD provides that the exception applies only to reproductions and extractions made 'for the purposes of text and data mining', while Article 4(2) permits copies to be retained only 'for as long as is necessary for the purposes of text and data mining'.
[173] '[...] the TDM output should not infringe any exclusive rights as it merely reports on the results of the TDM quantitative analysis, typically not including parts or extracts of the mined materials.' Geiger, Frosio, and Bulayenko (n 162).
[174] Okoń (n 112).
[175] Emanuilov and Margoni (n 26).
[176] Larroyed (n 98).

have limits. Two issues should be considered. First, whether the legislative and judicial intention can be understood as prioritising the development of technology to such an extent that it overrides the interpretation of copyright law. Second, whether refusing to apply the TDM exception to GenAI would make the provision itself ineffective.

The general principle of EU copyright law, that exceptions and limitations must be interpreted strictly within the limits of their purpose, while exclusive rights must be interpreted broadly in order to ensure a high level of protection for intellectual creations, implies that interpretative disputes should be resolved in *favorem auctoris.*[177] The CJEU has also emphasised that '[...] it precludes that provision [the exception] from being understood as requiring, beyond that limitation which is provided for expressly, copyright holders to tolerate infringements of their rights [...]'.[178]

At the same time, refusing to apply the TDM exception in cases of AI memorisation does not deprive the TDM exception of its proper effect. Article 4 of the DSM Directive was not designed specifically for GenAI, and there is even less indication that it was intended to ensure that the entire technical chain of every AI model is exempted from copyright liability. [179] This exception was designed to support data analysis activities in research and commercial contexts, where works are used as data sources to extract information. AI may be a tool for carrying out TDM, like other data analysis tools. However, if an AI system, in a specific case, goes beyond the function of information analysis and then continues to retain the work not for the purpose of continuing TDM, but in order to reproduce protected expression. Refusing to apply Article 4 to those further restricted acts is not an unreasonable narrowing of the exception, but rather, it preserves the internal boundary of the exception itself.

Accordingly, especially in the case of AI memorisation examined here, the exception cannot be extended to exempt the making of copies at the stage of AI training.

[177] Rendas (n 151) 93-94.
[178] Case C-435/12 *ACI Adam BV and Others v Stichting de Thuiskopie and Stichting Onderhandelingen Thuiskopie vergoeding*, EU:C:2014:254 [30]-[31].
[179]See also Eleonora Rosati, 'Copyright Exceptions and Fair Use Defences for AI Training Done for "Research" and "Learning," or the Inescapable Licensing Horizon' (2025) 16 European Journal of Risk Regulation 961.

In March 2026, the European Parliament adopted a resolution on Copyright and generative artificial intelligence, which reflects a relatively consistent view that the DSM Directive in particular, and copyright law more generally, are not sufficient to cover and effectively address the issues created by GenAI. [180] It therefore calls for amendment and the introduction of new regulatory mechanisms, especially for GenAI. On this basis, future debates may no longer focus mainly on the applicability of the TDM exception, but may shift towards the discussions of new policy measures.

4.3 Three-Step Test

In EU copyright law, exceptions must not only fall within the wording of a specific provision, but must also pass the three-step test as a final layer of review for the application of exceptions. Originating from Article 9(2) of the Berne Convention and later expanded in international instruments such as TRIPS and the WCT, the three-step test is now regarded as the 'centrepiece of the exceptions regimes' in international copyright law. [181] This standard was adopted by the EU from those international instruments and incorporated into Article 5(5) of the InfoSoc Directive, according to which:

> *'The exceptions and limitations [...] shall only be applied in certain special cases which do not conflict with a normal exploitation of the work or other subject-matter and do not unreasonably prejudice the legitimate interests of the rightholder.'*

In the EU copyright law context, the three-step test is not only a standard for legislators when designing or implementing exceptions, but also a limit for courts when interpreting and applying exceptions in individual cases. [182] This test reflects the nature of exceptions in copyright law: an act that interferes with an exclusive right can be justified as deserving exemption from legal liability only where that exemption is sufficiently legitimate and reflects a balance between the interests of users, the public interest, and the legitimate interests of authors or rightholders.

[180] European Parliament (n 165).
[181] Rendas (n 151) 97-98.
[182] Rendas (n 151) 104-111.

As analysed in the previous sections, existing exceptions such as temporary reproduction, incidental inclusion, and Article 4 of the DSM Directive are difficult to apply directly to GenAI memorisation. Therefore, this section does not repeat the assessment of whether each of those exceptions may apply. Instead, the underlying formula of Article 5(5) of the InfoSoc Directive will be used as a final evaluative framework to consider whether, given the nature of GenAI memorisation, this type of act can justify an exemption from copyright liability.

This section emphasises the risk of actual harm to the normal exploitation of the original works and to the legitimate interests of authors, as well as to both existing and potential markets.

One of the issues for which GenAI has been most criticised in society arises from the risk of harm to the creative labour market. Increasingly, reality shows that these are not merely vague or speculative concerns, but developments that have already occurred and have been demonstrated through market surveys and credible studies. [183]

GenAI is created and developed with the purpose of producing intellectual products similar to those created by humans, and the products it generates compete economically and directly with original authors. Importantly, these are not works created entirely independently by GenAI. They originate from the fact that GenAI has consumed and retained a massive amount of data from original works, and then uses it to reproduce products that are substantially similar or comparable and capable of acting as substitutes. This interference with exclusive rights not only creates substitute products, but also does so at extraordinary speed, on a massive scale, and at very low cost, or even close to zero cost in the case of some free versions of the model.

It is precisely these factors that allow GenAI-generated products to substitute for and reduce demand for human-created content, thereby taking market share away from

[183] See, for instance Maggie Harrison, 'AI Is Slitting the Throat of the Journalism Industry' (*Futurism*, 12 April 2024) <https://futurism.com/ai-google-discover-journalism-industry> accessed 25 May 2026; Alexandra Bruell, 'News Sites Are Getting Crushed by Google's New AI Tools' (*The Wall Street Journal*, 23 March 2024) <https://www.wsj.com/tech/ai/google-ai-news-publishers-7e687141> accessed 25 May 2026; Akila Quinio, '"They wanted me to make myself obsolete": translators find themselves at the sharp end of AI' *(Financial Times*, 20 June 2023) <https://www.ft.com/content/50b1f03e-1d10-4a7c-afa3-b48e9a8d5133> accessed 25 May 2026.

original works. This trend has already emerged and is becoming more serious as GenAI becomes more widespread and continues to develop.

Surveys in the translation industry show that more than one-third of translators have lost work due to GenAI and more than 40% have experienced a reduction in income. Labour market studies have also recorded signs of substitution between human workers and machines in writing and translation. A recent survey of verified professional artists shows that many reported fewer job opportunities and increasing professional pressure.[184] These forms of market harm have also been raised in claims in GenAI-related litigation. For example, in *The New York Times v Microsoft*, the claimant argued that chatbot outputs could divert traffic away from the newspaper's website, thereby affecting advertising revenue and the subscription model, that is, the publisher's normal channels of exploitation. [185] In *GEMA v OpenAI*, GEMA argued that OpenAI harmed the licensing framework for the use of musical works in both training and output.[186]

As regards the data licensing market more generally, this market is currently limited to some extent by the TDM exception for acts of data collection and analysis. Viewed in light of the three-step test, it becomes even clearer that the purpose of the TDM exception is to enable the creation of a different product serving a different and unrelated purpose, without harming the market for the exploitation of the original product. [187] If the entire process of GenAI training were allowed to benefit from an exception for free data collection, rightholders would suffer double harm: they would lose revenue from the licensing market that they would normally be entitled to, while also bearing the economic consequences of the results produced through GenAI's exploitation of their own works.

It may be argued that such market harm cannot be attributed to AI, because the model merely learns from existing works and generates new creative products, in a way that is similar to human learning. However, such a comparison is misleading and fails to

[184] Harry H. Jiang, Jordan Taylor, and William Agnew, 'How Professional Visual Artists are Negotiating Generative AI in the Workplace' (2024) <https://doi.org/10.48550/arXiv.2603.04537 >.
[185] *The NYT v Microsoft* (n 4).
[186] *GEMA* (n 7).
[187] Rossana Ducato and Alain Strowel, 'Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out' (2021) 43 European Intellectual Property Review 322. <https://ssrn.com/abstract=3829858>.

capture the real nature of the issue. GenAI cannot be equated with human creators by looking only at isolated acts of learning or creation. Its economic significance lies in the structural asymmetry of speed, scale and cost: AI systems can produce substitutive outputs in quantities and at a pace that human authors cannot realistically match. This changes the market effect of the activity and makes the harm to normal exploitation more serious than a simple comparison with human creative influence would suggest.
As a result, AI models may negatively affect interests that have traditionally belonged to rightholders and that authors could exploit through contracts, licensing, or publication. Therefore, allowing AI to benefit from an exception, whether through the TDM exception or a newly created exception, without establishing a mechanism to ensure reasonable compensation for rightholders, would be difficult to reconcile with the conditions required by the three-step test.

5 Responses and Possible Solutions: Technical and Legal Approaches

For this part, the analysis turns to the discussion on how such copies could be treated. This includes technical efforts to address memorisation, how the nature of the phenomenon could be understood, how responsibility could be allocated, how the interests of different stakeholders could be balanced, and what new policies could be developed in the near future for a latent, distributed, and difficult-to-retrieve form of copy within a technology which has now become widely used and deeply embedded in cultural and social life.

5.1 Technical Measures to Address AI Memorisation

Technical measures used to reduce AI memorisation can be divided into three groups, corresponding to three stages in the ‘lifecycle’ of AI operation: pre-training data, the training process, and output generation.

At the stage of data collection and preparation, studies propose measures such as deduplication, data sanitisation, or data curation. These involve removing duplicate data, near-duplicate data, or high-risk data, or building datasets that are better selected,

labelled, and controlled in terms of licensing. [188] From a technical perspective, deduplication aims to remove repeated or near-repeated data, because repeated exposure to the same sequence may increase the likelihood that the model memorises that sequence and reproduces it in the output. In addition, data sanitisation and data curation also help control the legal status of the data, including its source, licence conditions, attribution or notice requirements, and whether the rightholder has opted out or reserved rights, thereby reducing legal risks from the stage of constructing the training corpus.

During the training stage, methods such as differential privacy, regularisation, fine-tuning, or machine unlearning are used to make the model less likely to disclose memorised information, or better able to remove such information. However, recent studies still show different levels of effectiveness. Regularisation may be slow and less effective, while fine-tuning or unlearning appears more promising but remains costly and difficult to verify at scale. [189] In any event, completely eliminating memorisation remains impossible, because it is still a feature of machine learning models.

The debate on memorisation first emerged in the machine learning community around concerns that AI could leak personal data and infringe privacy, and later developed into allegations of copyright infringement at the output stage. For this reason, many AI companies are particularly concerned with reducing output-level infringement through measures such as output filtering, memory-free decoding, or code referencing, as used by GitHub for computer programs, in order to detect and block, or at least warn against, generated content that matches or nearly matches public data. [190] However, these filters only address the risk after the output has been generated. They depend on technical matching thresholds and may miss forms of memorisation that are not fully verbatim. Therefore, although AI providers continue to improve these filters, they still cannot prevent this phenomenon effectively.

[188] Emanuilov and Margoni (n 26) 25-27.
[189] Mansi Sakarvadia and others, 'Mitigating Memorization In Language Models' (2024) arXiv:2410.02159 <https://doi.org/10.48550/arXiv.2410.02159>.
[190] Ippolito and others (n 50).

The fact that an AI provider may prevent, reduce, or even completely remove infringing outputs does not resolve the problem that the copy may still remain inside the model. Output filters only intervene at the stage where protected expression is externally manifested, while the act of reproduction at the model level remains. Measures controlling the risk of harm at the output stage can therefore only be understood as a responsibility to reduce the further reproduction and dissemination of works to the public. They do not amount to a complete exemption for the earlier act of reproduction. The question is whether, if output-level infringement were fully resolved through highly advanced techniques in the future, reproduction inside the AI model would still be treated as infringement. In my view, the answer should not be negative. The reproduction right and the right of communication to the public are clearly established as two distinct rights under EU copyright law. If AI providers were allowed to avoid all liability simply by ensuring that the work is never 'regurgitated', the reproduction right would be reduced to a disguised right of control over communication. This would not be consistent with the principles of copyright law, where the reproduction right protects authors from the act of making a copy itself, not only from the moment when that copy is disclosed or disseminated.

If this situation is compared to private copying by individuals, it should be noted that even the private copying exception under Article 5(2)(b) of the InfoSoc is not treated as an absolute freedom to copy. The copy must not be made for directly or indirectly commercial purposes, and it is also linked to a mechanism of fair compensation for rightholders. Therefore, there is no reason to create a separate exception under which a copy made inside a model, and operating as part of the profit-generating system of a commercial AI provider, would automatically be treated as lawful merely because no further copy appears in the public output.

5.2 Balancing Evidentiary Burdens and Responsibility

Some scholars have proposed a mechanism of 'absolution' or of placing memorisation outside the reproduction right. [191] This solution is based on the argument that

[191] Eg Okoń (n 112), Emanuilov and Margoni (n 26).

memorisation, or more precisely ‘regurgitation’, does not always occur. A more flexible treatment may therefore be acceptable, because the data has been ‘decomposed’ to such an extent that it is no longer an identifiable ‘work’ within the parameters, and thus need not be treated as strictly as an ordinary copy. The mere existence of a copy should not automatically be treated as copyright infringement. Instead, liability should arise only at the output stage, when AI ‘regurgitates’ responses containing protected elements of original works. On this view, the responsibility of AI providers would be framed through duties of control, accountability, and reasonable care, together with ‘safe harbour’ provisions. This approach is supported by GenAI companies and commentators who favour the development of GenAI.

In the same context, one scientific article measuring AI memorisation relies on a counterfactual criterion: whether, without the existence of the work or group of works in the training data, the AI would still be able to generate valuable content.[192] This counterfactual idea is useful, but it should not be used in the way the authors seem to suggest. [193] The argument should be shifted from an individual causal relationship that each claimant must prove again from the beginning, towards showing a general premise about the inherent dependence of AI models on pre-existing works.

These points also open up new approaches to the burden of proof and the assessment of evidence in litigation. If the ‘ingestion’ of works into AI models is regarded as reproduction, and the causal link between GenAI outputs and pre-existing works is apparent, the claimant’s burden of proof may be reduced in two ways. First, where rightholders can rely on a relatively reasonable measurement method or output, they would not have to prove again whether the work was included in the input. Second, rightholders would not have to prove infringement at the output stage, but only that their work was included in the AI’s training data. The burden of proof could then shift to AI companies. On the assumption that an act of reproduction exists, they would have

[192] Annie Liang and Jay Lu, 'Creative Ownership in the Age of AI' (2026) arXiv:2602.12270 <https://doi.org/10.48550/arXiv.2602.12270>.

[193] The authors appear to advocate imposing an additional evidentiary requirement on claimants. Such an approach would place on plaintiffs an almost insurmountable burden of proof: they would have to demonstrate that a specific AI output could not have been generated without the inclusion of their particular work in the training dataset. In the context of GenAI models trained on enormous datasets, this approach is impractical and does not reflect the direction currently taken by courts and lawmakers.

to prove either that no copying occurred at the model level or that the work was not present in the input data.

This approach greatly benefits rightholders and provides them with a strong protection mechanism. Its weakness, however, is that it seems to merge all technical issues of AI into a single category. If this approach were actually applied, almost all GenAI training on copyrighted material would become an infringement risk. To some extent, that may be true. More importantly, however, it could lead to negative consequences. It risks encouraging arbitrary claims by rightholders, and gradually pushing the understanding of this concept away from its real nature, thereby distorting future legislative and judicial responses to GenAI. [194] This could create an imbalance between protecting the interests of rightholders and promoting technological development.

A more balanced approach, which is also supported by many current views and is closer to the reasoning of the Munich court, is to recognise that memorisation may have legal relevance, without equating every case of 'ingestion' with reproduction. A full evidential burden should still be required in each individual case, and the evidence of retrieval must be sufficiently convincing. In practice, however, future cases will not all be the same. Proving the existence of copies of works inside a model may become more difficult, especially where AI companies apply output filters. [195]

It is necessary to take into account the disadvantaged position of rightholders in accessing evidence, especially where the rightholders are individuals, ordinary groups of authors, or small organisations without sufficient resources and technical tools. If the mechanism places the entire burden of proof on the claimant, rightholders may be unable to protect themselves against copyright infringements that clearly occur, simply because they lack access and control. They have no access to the system or to source tracing. They cannot observe billions of private interactions between users and chatbots, check every output, or know in advance when, in what context, and with which user the model will reproduce their protected expression in responses delivered to hundreds of

[194] Larroyed (n 98).
[195] von Lewinski (n 148).

millions of users every day. [196] They may not even know whether their works have been incorporated into AI models, and it is difficult for them to prove this without technical assistance or an admission from the AI company. Cases involving AI-generated or AI-assisted productions by users that are accused of plagiarism are only the surface manifestation of a deeper problem that creative elements have been absorbed into the system in a way that is extremely difficult to trace and control.

The instability of output retrieval may be used to oppose absolute liability at the model level, but this argument can also turn back against AI companies. That instability makes it difficult for both sides to prove either the existence or the absence of reproduction. AI providers are the parties who are fully aware that they control a high-risk system that may contain copies and have an inherent capacity to reproduce them, but they also cannot fully control or guarantee that those AI models will not infringe.

In this context, the recommendation of the Danish Expert Group is particularly useful. It links a presumption to a specific transparency obligation concerning training data. Accordingly, if an AI provider does not provide datasets or relevant information to rightholders within a reasonable time, a reversal of the burden of proof would be triggered. The rightholder's content would then be presumed to have been used for training, unless the provider proves otherwise. [197] This would lead to compliance and risk-governance obligations for AI companies, such as ex ante obligations relating to documentation, transparency, testing for memorisation, deduplication, compliance with reservations of rights, and duties to explain and disclose technical audits, training sources, or certain aspects of model operation. Such obligations are also consistent with the risk-control principles under the AI Act. This mechanism does not completely simplify the rightholder's burden of proof. The rightholder would still need sufficiently strong initial evidence and arguments, but part of the evidential burden would be shifted to AI providers.

[196] Numerous reports and statistics have already shown the enormous and rapidly growing number of daily GenAI users, for instance, see Aaron Chatterji and others, 'How People Use ChatGPT' (2025) NBER Working Paper No w34255 <https://ssrn.com/abstract=5487080> accessed 24 May 2026; OpenAI, 'The State of Enterprise AI' (*Open AI*, 8 December 2025) <https://openai.com/vi-VN/index/the-state-of-enterprise-ai-2025-report/>accessed 24 May 2026.

[197] Expert Group on Copyright and Artificial Intelligence, *Recommendations from Expert Group on Copyright and Artificial Intelligence* (Danish Ministry for Culture, 2025) 57.

5.3 Fair Remuneration and Future Directions

Looking at the broader picture, it is necessary to consider the larger question of the value gap: whether it is fair for AI providers to commercialise value built from the collection of massive amounts of copyrighted works, while legal responsibility only truly arises when a specific infringement is detected and fully proved. If so, much of the economic value and technological advantage generated from the use of works at the input stage risks falling outside the control of rightholders. Therefore, there could be a mechanism that gives rightholders control, or at least appropriate remuneration, for the use of their works before that value is transformed by the model and exploited in the market. This is consistent with the logic of building copyright protection for works against infringement risks in the digital environment, where a mechanism giving rightholders control from the input stage is 'a suitable basis for the regulation [...] of problems arising from the computer storage and retrieval of protected works'.[198]

The mechanism most often proposed in this context is licensing. Licensing allows rightholders to control the use of their works, while also setting the scope of use, the purpose of use, attribution obligations, transparency requirements, remuneration, or limits on downstream use. Collective licensing and extended collective licensing have received particular attention of commentators, content creators, and policy discussions as possible responses.[199] These mechanisms assist individual authors and small creators, who often lack the bargaining power, monitoring capacity, or resources to enforce their rights against large AI providers. At the same time, they also reduce transaction costs for AI developers, because they do not need to seek permission from each individual rightholder separately.[200]

---

[198] See WIPO, 'Problems arising from the use of electronic computers and other technological equipment'(1972) 8(1) Copyright, reporting Professor Eugen Ulmer's view that the Berne Convention protected works at the input stage and that input into a computer should be considered reproduction.

[199] See Lucchi (n 169); European Parliament (n 165); Giancarlo Frosio, 'Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI – Part 2' (*Kluwer Copyright Blog*, 11 December 2025) < https://legalblogs.wolterskluwer.com/copyright-blog/copyright-in-formaldehyde-how-gema-v-openai-freezes-doctrine-and-chills-ai-part-2/> accessed 24 May 2026; Johan Axhamn, 'Extended Collective Licensing for Use of Copyrighted Works for Machine Learning' (2025) 48(4) The Columbia Journal of Law & the Arts 523.

[200] Cf Stepanka Havlikova, 'Collective Licensing for Gen AI Training: Feasible or Flawed? – Part 2' (Kluwer Copyright Blog, 20 May 2026) <https://legalblogs.wolterskluwer.com/copyright-blog/collective-licensing-for-gen-ai-training-feasible-or-flawed-part-2/> accessed 24 May 2026, discussing structural difficulties of collective and extended collective licensing for GenAI training.

Alongside licensing, other remuneration mechanisms have also been proposed, such as a statutory remuneration right or an AI levy, under which AI providers would be required to pay fair remuneration to authors, potentially through mandatory collective management or social and cultural funds. [201]

Nevertheless, any fee-based or licensing mechanism still needs to be designed carefully. Its aim could not be merely to create an additional revenue stream for large corporations in the creative industries, but should pay particular attention to the interests of individual authors, those who directly create creative value. Otherwise, licensing may simply redistribute benefits between large businesses: on one side, AI companies, and on the other, companies controlling the content industries, while authors remain in a disadvantaged position.

Therefore, this mechanism needs to be accompanied by strict rules on compliance obligations, transparency, benefit allocation, and monitoring, in order to ensure that the recovered value actually reaches creators. This should also be supported by trusted intermediary institutions that help standardise, record, and verify rightholders' rights. The European Parliament's proposal to entrust EUIPO with supporting licensing processes, managing or listing exclusions of copyrighted works from AI training, and raising awareness partly reflects the direction in which the EU may move towards building such an institutional infrastructure. [202]

As regards acts of data collection and analysis at the pre-training stage, many views suggest that the text and data mining exception under Article 4 of the DSM Directive has not created a sufficiently balanced mechanism for the problems raised by GenAI. After three years of facing challenges from GenAI, the opt-out mechanism under Article 4 has revealed many practical limitations, as rightholders have not been able to exercise this right effectively. The position of the European Parliament also reflects a relatively clear view that this exception was not drafted with the purpose of allowing

[201] Martin Senftleben, 'Generative AI and Author Remuneration' (2023) 54 IIC 1535 DOI: 10.1007/s40319-023-01399-4; Kaigeng Li, Hong Wu and Yupeng Dong, 'Copyright Protection During the Training Stage of Generative AI: Industry-Oriented U.S. Law, Rights-Oriented EU Law, and Fair Remuneration Rights for Generative AI Training Under the UN's International Governance Regime for AI' (2024) CLSR 106056 DOI: 10.1016/j.clsr.2024.106056.
[202] European Parliament (n 165) 13.

the mass use of copyright-protected materials by GenAI, especially where such use leads to the creation of a competing product made available to the public. It also directly criticises the fact that the EU legislator incorporated Article 4 of the DSM Directive into the AI Act without clarifying the legal consequences of relying on this exception in the GenAI context. According to the report, the proposed solution is to combine licensing possibilities with transparency requirements and a more effective opt-out mechanism, one that is standardised, machine-readable, and capable of being recorded in a European register. [203]

A stronger direction was proposed by the Danish Expert Group in the report prepared at the request of the Danish Ministry of Culture. In addition to recommending that the EU adopt harmonised rules to revise the opt-out mechanism, so that rightholders can use this right effectively in practice, the Expert Group also appears to lean towards moving from an opt-out model to an opt-in model if these improvements do not produce results. This issue, in its view, should be addressed quickly in 2026. [204]

In addition, an appropriate solution does not necessarily have to be an absolute and uniform obligation for all GenAI systems or all providers. It may instead be designed in a tiered way, based on differences in scale, commercial purpose, the risk level of the model, the type of data used, and the compliance capacity of each actor. Developers of large-scale foundation models or GPAI systems, especially commercial companies with significant market exploitation capacity, may be subject to stricter obligations. By contrast, SMEs, research organisations, or non-commercial models may need more flexible mechanisms in order to develop.

6 Final Remarks and Insights

This thesis begins with a question that whether memorisation in Generative AI models, particularly LLMs, can amount to reproduction of protected works under EU copyright law. On the basis of the analysis developed above, it is argued that AI memorisation can, in principle, fall within the scope of the reproduction right.

[203] European Parliament (n 165) 19.
[204] Expert Group on Copyright and Artificial Intelligence (n 197) 59.

To reach this conclusion, the thesis first clarified that memorisation is no longer a merely speculative concern, but a phenomenon that can be observed and demonstrated through technical studies. On that basis, the thesis analysed the reproduction right in EU copyright law as a broad, functional right grounded in technological neutrality. Although protected expression does not exist in model parameters in the same way as in an ordinary digital file, it can still be regarded as a copy where there is sufficiently convincing evidence that the protected expression has been retained at the model level. However, this approach does not mean that the entire model, the entire training dataset, or every statistical pattern must be treated as a copy under copyright law.

The thesis then considered whether existing exceptions and limitations could justify this form of reproduction. The analysis showed that temporary reproduction, incidental inclusion and Article 4 of the DSM Directive are all unlikely to cover cases of GenAI memorisation where protected expression is retained and remains capable of being reproduced later. If the law wishes to design a new exception or exemption mechanism for this phenomenon, such a mechanism must take into account fair remuneration and the legitimate interests of rightholders.

Therefore, the remaining issue is not only whether memorisation constitutes reproduction, but how the law could treat this new form of reproduction. This thesis does not seek to cover all emerging judicial and policy responses, but offers several reflections on important directions that could be considered. At present, many AI providers still rely mainly on output filters to reduce ‘regurgitation’. However, this is a remedy that addresses the symptoms rather than the root cause: it intervenes only after the protected expression has already been reproduced or retained within the system, therefore does not directly address the ‘reproduction inside the model’ problem. Moreover, in practice, these filters still cannot guarantee full and consistent effectiveness.

Recognising AI memorisation as reproduction has value beyond merely redefining the term reproduction or proving legal liability in individual cases. It exposes a structural dependency that is often obscured by technical descriptions of training. Once a model can reproduce or nearly reconstruct protected expression, the argument that training is only an abstract process of learning can no longer stand. Memorisation forces copyright

law to confront more directly the fact that GenAI systems are built on, and continue to require, human creative input.

This leads to a broader policy consequence, where AI development cannot be treated as if it were independent of the creative labour it consumes. Creators must retain meaningful control over the use of their works at the input stage and receive fair remuneration when those works sustain the AI ecosystem.

Of course, these measures have faced, and will continue to face, objections not only from AI companies but also from perspectives that prioritise technological development and economic competitiveness. Some may worry that placing too much emphasis on the storage and exploitation of works inside AI systems would interpret the law in a way that over-protects copyright holders. Others may argue that strict compliance duties, permission requirements or large-scale payment obligations could create enormous transaction costs, slow down AI development, reduce the competitiveness of EU businesses, and even encourage AI companies to relocate to more permissive jurisdictions. This is a concern that cannot be ignored, especially in a context where AI has become a strategic field of competition between major economic powers.[205]

When the costs borne by technology companies are placed at the centre of the discussion, another type of cost is often overlooked or understated. When artists lose creative contracts, suffer reduced income, lose bargaining power, are displaced in the market, lose professional opportunities, and face a decline in the incentive for human creativity, these are costs as well. The question is whether these costs should be borne by the companies that commercially benefit from the large-scale use of protected works, or by the creative communities whose works make these systems possible in the first place.

From the perspective of fair cost allocation, where an activity extracts value from a shared social resource, the burden of a socially transformative development should

[205] See European Commission, 'The EU invests in artificial intelligence only 4% of what the U.S. spends on it' (European Commission, 11 January 2025) <https://ec.europa.eu/newsroom/eismea/items/864247/en> accessed 24 May 2026.

primarily fall on those who have extracted and captured the greatest value from it, or, as some would put it, ‘tax the rich’.

From a long-term perspective on social development, the sustainability of AI innovation cannot be separated from the sustainability of human creativity. This is not only a moral or cultural concern, but also a technical one. GenAI is often described as a machine that is constantly hungry for new data. This is not merely a metaphor: technically, its performance depends on the continuous availability of human-created materials at scale, with sufficient diversity and quality. If future models are trained mainly on outputs generated by earlier models, the system risks falling into recursive degradation. [206]

The availability of cheap and instant machine-generated content may satisfy short-term market demand, but it does not by itself secure the conditions for cultural renewal. It was said that a market flooded with synthetic outputs may increase the quantity of content while weakening the economic and social basis for future human creativity.[207] To put it simply, if human creative labour is gradually exhausted and degraded, that may also be the point at which GenAI begins to collapse because it no longer has its ‘lifeblood’. By then, restoring the creative ecosystem to its original condition would not be easy. That is a cost no society or economy can afford to pay.

Human creativity should not be treated as something that can be temporarily sacrificed in favour of growth figures on paper. It is part of the social and cultural infrastructure through which communities preserve memory, express identity, transmit values and carry lasting influence into the future of humanity. For that reason, AI technology, like

[206] See for instance Ilia Shumailo and others, ‘The Curse of Recursion: Training on Generated Data Makes Models Forget’ (2023) arXiv:2305.17493v3 <https://doi.org/10.48550/arXiv.2305.17493>; Matthias Gerstgrasser and others, ‘Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data’ (2024) arXiv:2404.01413v2 <https://doi.org/10.48550/arXiv.2404.01413> ; Ilia Shumailov and others, ‘AI Models Collapse When Trained on Recursively Generated Data’ (2024) 631 Nature 755 DOI: 10.1038/s41586-024-07566-y.

[207] In *Kadrey v Meta*, Judge Chhabria repeatedly emphasised this concern, stating in particular that ‘by training generative AI models with copyrighted works, companies are creating something that [...] dramatically undermine the incentive for human beings to create things the old-fashioned way’; and that ‘copying the protected works, however transformative, involves the creation of a product with the ability to severely harm the market for the works being copied, and thus severely undermine the incentive for human beings to create.’ *Kadrey* (n 5) 2-3.

any other technology, exists to serve the cultural sovereignty of humanity, not the other way around.[208]

Copyright is, by nature, the outcome of a negotiation between different interest groups in search of a point of balance,[209] and that outcome inevitably reflects which values are being prioritised. Yet, regardless of the balance ultimately adopted, a policy that accepts the weakening of the human creative ecosystem in order to make AI development cheaper risks 'eating the seeds of the next harvest' in both cultural and technological production. Encouragingly, this concern is not raised only by cultural or intellectual property organisations, but is also reflected in recent legislative and policy signals at both EU and Member State level, including the European Parliament's 2026 Resolution on copyright and generative AI.[210]

At present, much attention is directed towards the future preliminary ruling of the CJEU in Case C-250/25 *Like Company v Google Ireland*, where the Court is expected to provide, for the first time, guidance on the relationship between GenAI, copyright and related rights under EU law. At the same time, the next steps of EU lawmakers in developing and revising the legal framework are also being closely watched. Whether EU copyright law can adapt to the age of GenAI without sacrificing the creative values it was designed to protect will remain one of the central questions of the years ahead.

[208] UNESCO, 'A new expert report explores how AI is transforming culture' (*UNESCO*, 25 November 2025) <https://www.unesco.org/en/articles/new-expert-report-explores-how-ai-transforming-culture> accessed 24 May 2026.
[209] Blayne Haggart, *Copyfight: The Global Politics of Digital Copyright Reform* (University of Toronto Press 2014) 22.
[210] European Parliament (n 165).

Bibliography

Treaties and Legislation

Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971)

International Convention for the Protection of Performers, Producers of Phonograms and Broadcasting Organisations (Rome Convention) 1961

WIPO Copyright Treaty (WCT) 1996

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence

Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society

Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market

Cases law

CJEU Case law

Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening* EU:C:2009:465

Case C-5/08 *Infopaq International A/S v Danske Dagblades Forening*, Opinion of AG Trstenjak, EU:C:2009:89.

Joined Cases C-403/08 and C-429/08 *Football Association Premier League Ltd and Others v QC Leisure and Others and Karen Murphy v Media Protection Services Ltd* EU:C:2011:631

Case C-145/10 *Eva-Maria Painer v Standard VerlagsGmbH and Others* EU:C:2011:798

Case C-302/10 *Infopaq International A/S v Danske Dagblades Forening* EU:C:2012:9

Case C-435/12 *ACI Adam BV and Others v Stichting de Thuiskopie and Stichting Onderhandelingen Thuiskopie vergoeding*, EU:C:2014:254

Case C-360/13 *Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd* EU:C:2014:1195

Case C-310/17 *Levola Hengelo BV v Smilde Foods BV* EU:C:2018:899

Case C-476/17 *Pelham GmbH v Ralf Hütter* EU:C:2019:624

Case C-250/25 *Like Company v Google Ireland Ltd* (pending)

German case law

GEMA v OpenAI No 42 O 14139/24 (LG München I 11 November 2025)

Laras Tochter, Case No I ZR 65/96 (Bundesgerichtshof, 29 April 1999), BGHZ 141, 267

Möbelkatalog, Case No I ZR 177/13 (Bundesgerichtshof, 17 November 2014)

Robert Kneschke v. LAION e.V., Case No. 310 O 227/23 (Hamburg Regional Court 27 September 2024)

UK Case law

Getty Images (US) Inc v Stability AI Ltd [2025] EWHC 2863, [2026] Bus LR 582, [2025] WLR(D) 571

US Case law

Authors Guild v OpenAI Inc No 1:23-cv-08292 (SDNY 19 September 2023)

Bartz v Anthropic PBC No 3:24-cv-05417 (ND Cal 19 August 2024)

Carreyrou v Anthropic PBC No 3:25-cv-10897 (ND Cal 22 December 2025)

Denial v OpenAI Inc No 3:25-cv-05495 (ND Cal 24 July 2025)

Doe 1 v GitHub Inc No 4:2022cv06823 (ND Cal 22 January 2024)

Kadrey v Meta Platforms Inc No 3:23-cv-03417 (ND Cal 7 July 2023)

Nichols v Universal Pictures Corp 45 F 2d 119 (2d Cir 1930)

Re OpenAI Inc Copyright Infringement Litigation 1:25-md-03143 (SDNY 11 April 2025)

The New York Times Co v Microsoft Corp No 1:23-cv-11195 (SDNY 27 December 2023

The New York Times Co v Perplexity AI Inc No 1:25-cv-10106 (SDNY 21 October 2024)

EU Institutional Documents

European Commission, ‘Principles and guidelines for the Community's audiovisual policy in the digital age’ (Communication) COM (1999) 657 final

European Commission, ‘Proposal for a Directive of the European Parliament and of the Council on copyright in the Digital Single Market’ COM (2016) 593 final

European Parliament, ‘Amendments adopted by the European Parliament on 12 September 2018 on the proposal for a directive of the European Parliament and of the Council on copyright in the Digital Single Market’ P8_TA0337 (12 September 2018)

European Parliament, European Parliament resolution of 10 March 2026 on copyright and generative artificial intelligence – opportunities and challenges (2025/2058(INI))

Brando A, *Technological Aspects of Generative AI in the Context of Copyright, Attribution and Novelty in Generative AI Hypersurfaces* (Requested by the JURI Committee, Policy Department for Justice, Civil Liberties and Institutional Affairs Directorate-General for Citizens' Rights, Justice and Institutional Affairs PE 776.529, July 2025)

Geiger G, Frosio G and Bulayenko O, *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects* (In-Depth Analysis requested by the JURI Committee, European Parliament Policy Department for Citizens' Rights and Constitutional Affairs, PE 604.941, 2018)

Lucchi N, *Generative AI and Copyright: Training, Creation, Regulation* (Study requested by the JURI Committee, Policy Department for Justice, Civil Liberties and Institutional Affairs Directorate-General for Citizens' Rights, Justice and Institutional Affairs, PE 774.095, July 2025)

Madiega T, 'Copyright in the Digital Single Market' (Briefing, EPRS, PE 593.564, 2019)

EUIPO, *The Development of Generative Artificial Intelligence from a Copyright Perspective* (EUIPO 2025)

International Organisation Documents

Expert Group on Copyright and Artificial Intelligence, *Recommendations from Expert Group on Copyright and Artificial Intelligence* (Danish Ministry for Culture 2025)

WIPO, 'Draft model provisions for legislation in the field of Copyright, III. Comments on the draft model provisions for legislation in the field of copyright' (1989) Document CE/MPC I/2-III 8

WIPO, *Records of the intellectual property conference of Stockholm* (WIPO Publication No 311, 1971)

Books

Al-Kaswan A, Izadi M and van Deursen A, ‘Traces of Memorisation in Large Language Models for Code’ in *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering* (ACM 2024)

Carlini N and others, 'Extracting Training Data from Diffusion Models' in *Proceedings of the 32nd USENIX Conference on Security Symposium* (USENIX Association, 2023)

Carlini N and others, ‘Extracting Training Data from Large Language Models’ in *Proceedings of the 30th USENIX Security Symposium* (USENIX Association 2021)

Ippolito D and others, ‘Preventing Generation of Verbatim Memorization in Language Models Gives a False Sense of Privacy’ in *Proceedings of the 16th International Natural Language Generation Conference* (Association for Computational Linguistics 2023)

Feldman V, 'Does learning require memorization? A short tale about a long tail' in Konstantin Makarychev and others (eds), *STOC ’20 Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing* (ACM, 2020)

Ficsor M, *The Law of Copyright and the Internet: The 1996 WIPO Treaties, Their Interpretation and Implementation* (Oxford University Press 2002)

Ginsburg JC and Treppoz E, *International Copyright Law: U.S. and E.U. Perspectives: Text and Cases* (Edward Elgar 2015)

Guibault L, *Copyright Limitations and Contracts* (Dalhousie University Schulich School of Law 2002)

Haggart B, *Copyfight: The Global Politics of Digital Copyright Reform* (University of Toronto Press 2014)

Koh PW and Liang P ‘Understanding black-box predictions via influence functions’ in Precup d and Teh Y W(eds), *Proceedings of the 34th International Conference on Machine Learning* (JMLR.org 2017) vol 70

Masouyé C, *Guide to the Berne Convention for the Protection of Literary and Artistic Works (Paris Act, 1971)* (WIPO 1978),

Mueller FB and others, 'LLMs and Memorization: On Quality and Specificity of Copyright Compliance' in Das S and others (eds), *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society* (AIES-24) (AAAI Press, 2024) vol 7, no 1

Rendas T, *Exceptions in EU Copyright Law: In Search of a Balance Between Flexibility and Legal Certainty* (Kluwer Law International BV 2021),

Smits JM, ‘What Is Legal Doctrine? On the Aims and Methods of Legal-Dogmatic Research’ in van Gestel R and others (eds), *Rethinking Legal Scholarship: A Transatlantic Dialogue* (Cambridge University Press 2017

Zhu L, Liu Z, and Han S, ‘Deep Leakage from Gradients’ in Hanna M Wallach and others (eds), Proceedings of the 33rd International Conference on Neural Information Processing Systems (Curran Associates Inc 2019

Journals

Axhamn J, ‘Extended Collective Licensing for Use of Copyrighted Works for Machine Learning’ (2025) 48(4) The Columbia Journal of Law & the Arts *523*

Cooper AF and Grimmelmann J, ‘The Files are in the Computer: On Copyright, Memorization, and Generative AI’ (2025) 100(1) Chi-Kent L Rev 141

Craig CJ, ‘Technological Neutrality: Recalibrating Copyright in the Information Age’ (2016) 17(2) Theoretical Inquiries in Law 601

Ducato S and Strowel A, ‘Ensuring Text and Data Mining: Remaining Issues with the EU Copyright Exceptions and Possible Ways Out’ (2021) 43 European Intellectual Property Review 322

GRUR, ‘Copyright Infringement in Form of a Reproduction of Preexisting Works in a Large Language Model’ (2026) 75(4) GRUR International 371

Jütte BJ and Quintais JP, ‘The Pelham Chronicles: Sampling, Copyright and Fundamental Rights’ (2021) 16 JIPLP 213

Leistner M and Antoine L, ‘TDM and AI Training in the European Union - From “LAION” to Possible Ways Ahead?’ (2025) 11 GRUR International 1032

Li K, Wu H and Dong Y, ‘Copyright Protection During the Training Stage of Generative AI: Industry-Oriented U.S. Law, Rights-Oriented EU Law, and Fair Remuneration Rights for Generative AI Training Under the UN's International Governance Regime for AI’ (2024) CLSR 106056

Lindberg M, ‘Applying Current Copyright Law to Artificial Intelligence Image Generators in the Context of Anderson v. Stability AI, Ltd’ (2024) 15 Cybaris 37

Lindberg V, ‘Building and Using Generative Models under US Copyright Law’ (2023) 18 Rutgers Bus L Rev 1

Matulionyte R, ‘Reconceptualising the Reproduction Right in the Age of AI’ (2025) 56 IIC 1914,

Mezei P, ‘A comprehensive guide to the InfoSoc Directive’ (2020) 15(1) JIPLP 70

Michels JD and Millard C, 'The New Things: Property Rights in Digital Files?' (2022) 81(2) The Cambridge Law Journal 323

Radeisen A, ‘Open Foundation Models and TDM Exceptions to Copyright – Building Blocks for an AI Ecosystem’(2026) 75 GRUR International 224

Reed C, ‘Online and Offline Equivalence: Aspiration and Achievement' (2010) 18 IJLIT 248

Rosati E, ‘Copyright Exceptions and Fair Use Defences for AI Training Done for “Research” and “Learning,” or the Inescapable Licensing Horizon’ (2025) 16 European Journal of Risk Regulation 961

Saw CL and Tan BZY, ‘Unpacking Copyright Infringement Issues in the GenAI Development Lifecycle and a Peek into the Future’ (2025) 58 CLSR 106163

Senftleben M, ‘Generative AI and Author Remuneration’ (2023) 54 IIC 153

Shilov I, Meeus M and de Montjoye Y-A, 'The mosaic memory of large language models' (2026) 17 Nat Commun 2142

Shumailov I and others, ‘AI Models Collapse When Trained on Recursively Generated Data’ (2024) 631 Nature 755

von Lewinski S, ‘GEMA v. OpenAI: The Munich Regional Court Presents an Exemplary Judgment on a Well Conceived Test Case’ (2026) 75(4) GRUR Int 340

WIPO, ‘Copyright’ (1982) Copyright 18(9)

WIPO, ‘Problems arising from the use of electronic computers and other technological equipment’ (1972) 8(1) Copyright

Online Sources and Working Papers

Ahmed A and others, ‘Extracting books from production language models’ (2026) arXiv:2601.02671 https://doi.org/10.48550/arXiv.2601.02671

Akdemir B, ‘Meta, Google & NVIDIA’s Landmark Study: When LLMs Stop Memorizing and Starts Generalizing’ (*Medium*, 11 June 2025) https://medium.com/@akdemir_bahadir/how-much-do-language-models-actually-remember-a-deep-dive-into-ai-memory-and-capacity-f7a71e55a93e

Bannigan MK and others, ‘AI Intellectual Property Disputes: The Year in Review and What’s Ahead’ (*Debevoise & Plimpton*, 2 December 2025) https://www.debevoise.com/insights/publications/2025/12/ai-intellectual-property-disputes-the-year-in

Bruell A, ‘News Sites Are Getting Crushed by Google’s New AI Tools’ (*The Wall Street Journal*, 23 March 2024) https://www.wsj.com/tech/ai/google-ai-news-publishers-7e687141

Carlini N and others, ‘Quantifying Memorization Across Neural Language Models’ (ICLR, Rwanda, 2023)

Chatterji A and others, ‘How People Use ChatGPT’ (2025) NBER Working Paper No w34255 https://ssrn.com/abstract=5487080

Cooper AF and others, 'Extracting memorized pieces of (copyrighted) books from open-weight language models' (SSRN, 2025) http://dx.doi.org/10.2139/ssrn.5262084

Cowen T, ‘Toothpick Producers Violate NYT Copyright’ *(Marginal Revolution*, 30 December 2023) https://marginalrevolution.com/marginalrevolution/2023/12/toothpick-producers-violate-nyt-copyright.html

Elangovan A, He J, and Verspoor K, ‘Memorization vs. Generalization: Quantifying Data Leakage in NLP Performance Evaluation’ (arXiv, 2021) https://doi.org/10.48550/arXiv.2102.01818

Emanuilov I and Margoni T, ‘Forget Me Not: Memorisation in Generative Sequence Models Trained on Open Source Licensed Code’ (SSRN, 2024) http://dx.doi.org/10.2139/ssrn.4720990

Emanuilov I and Margoni T, ‘Memorisation in Generative Models and EU Copyright Law: An Interdisciplinary View’ (*Kluwer Copyright Blog*, 26 March 2024) https://legalblogs.wolterskluwer.com/copyright-blog/memorisation-in-generative-models-and-eu-copyright-law-an-interdisciplinary-view

European Commission, ‘The EU invests in artificial intelligence only 4% of what the U.S. spends on it’ (*European Commission*, 11 January 2025) https://ec.europa.eu/newsroom/eismea/items/864247/en

Frosio G, ‘Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI – Part 2’ (*Kluwer Copyright Blog*, 11 December 2025)

https://legalblogs.wolterskluwer.com/copyright-blog/copyright-in-formaldehyde-how-gema-v-openai-freezes-doctrine-and-chills-ai-part-2/

Gao L and others, 'The Pile: An 800GB Dataset of Diverse Text for Language Modeling' (2020) arXiv:2101.00027 https://doi.org/10.48550/arXiv.2101.00027

Gesmer L, 'Copyright and the Challenge of Large Language Models (Part 3)' (*Mass Law Blog*, 27 November 2024) https://masslawblog.com/copyright/copyright-and-the-challenge-of-large-language-models-part-3/

GitHub, 'Finding Matching Code' (*GitHub Docs*) https://docs.github.com/en/copilot/how-tos/get-code-suggestions/find-matching-code

GitHub, 'Frequently asked questions: General: Does GitHub Copilot "copy/paste"?' (*GitHub*) https://github.com/features/copilot

Harrison M, 'AI Is Slitting the Throat of the Journalism Industry' (*Futurism*, 12 April 2024) https://futurism.com/ai-google-discover-journalism-industry

Hartill T and others, 'Do Smaller Language Models Answer Contextualised Questions Through Memorisation Or Generalisation?' (2023) https://doi.org/10.48550/arXiv.2311.12337

Havlikova S, 'Collective Licensing for Gen AI Training: Feasible or Flawed? – Part 2' (*Kluwer Copyright Blog*, 20 May 2026) https://legalblogs.wolterskluwer.com/copyright-blog/collective-licensing-for-gen-ai-training-feasible-or-flawed-part-2/

Shumailo I and others, 'The Curse of Recursion: Training on Generated Data Makes Models Forget' (2023) arXiv:2305.17493v3 https://doi.org/10.48550/arXiv.2305.17493

Jiang HH, Taylor J and Agnew W, 'How Professional Visual Artists are Negotiating Generative AI in the Workplace' (arXiv, 2024) arXiv:2603.04537 https://doi.org/10.1145/3772363.3799003

Jurafsky D and Martin JH, 'Speech and Language Processing' (2026) 3rd draft edn https://web.stanford.edu/~jurafsky/slp3

KODA, 'Historic ruling: OpenAI found guilty of exploiting music' (*KODA*, 11 of November 2025) https://koda.dk/en/about-koda/news/historic-ruling-openai-found-guilty-of-exploiting-music

Kuschel L and Rostam D, 'If It Looks Like a Duck On the Munich Regional Court's ruling in GEMA v. Open AI' (*Verfassungsblog*, 19 November 2025) https://verfassungsblog.de/munich-regional-courts-ruling-in-gema-v-open-ai/

Larroyed A, 'The Fallacy of the File: How the Memorisation Metaphor Misguides Copyright Law and Stifles AI Innovation' (SSRN 2025) http://dx.doi.org/10.2139/ssrn.5782882

Lee TB, 'Meta's Llama 3.1 can recall 42 percent of the first Harry Potter book' (*Understanding AI*, 12 June 2025) https://www.understandingai.org/p/metas-llama-31-can-recall-42-percent

Liang A and Lu J, 'Creative Ownership in the Age of AI' (2026) arXiv:2602.12270 https://doi.org/10.48550/arXiv.2602.12270

Maslej N and others, 'The AI Index 2025 Annual Report' (*Stanford University HAI*, April 2025) https://hai.stanford.edu/ai-index/2025-ai-index-report

Matthias Gerstgrasser and others, 'Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data' arXiv, (2024) arXiv:2404.01413v2 https://doi.org/10.48550/arXiv.2404.01413

Mezei P, 'Memorization and Generative AI - A Persistent Issue with Copyright Consequences?' (SSRN 2025) http://dx.doi.org/10.2139/ssrn.5404367

Nasr M and others, 'Scalable Extraction of Training Data from (Production) Language Models' (arXiv, 2023) arXiv:2311.17035, 15 https://doi.org/10.48550/arXiv.2311.17035

Nelson Elhage and others, 'Softmax Linear Units' (*Transformer Circuits Thread*, 27 June 2022) https://transformer-circuits.pub/2022/solu/index.html

Okoń Z, 'Memorization of Copyrighted Works by AI Models Under EU Law' (SSRN, 2026) http://dx.doi.org/10.2139/ssrn.6041395

OpenAI, ‘OpenAI and Journalism’ (*OpenAI*, 8 January 2024) https://openai.com/index/openai-and-journalism/

OpenAI, ‘The State of Enterprise AI’ (*Open AI*, 8 December 2025) https://openai.com/vi-VN/index/the-state-of-enterprise-ai-2025-report

Panwar M and others, 'For Better or for Worse, Transformers Seek Patterns for Memorization' (*NeurIPS*, 3 December 2025) https://neurips.cc/virtual/2025/loc/san-diego/poster/119570

Power A and others, ‘Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets’ (arXiv, 2022) http://arxiv.org/abs/2201.02177

Quinio A, ‘“They wanted me to make myself obsolete”: translators find themselves at the sharp end of AI’ (*Financial Times*, 20 June 2023) https://www.ft.com/content/50b1f03e-1d10-4a7c-afa3-b48e9a8d5133

Rando J and Tramèr F ‘Extracting (Even More) Training Data From Production Language Models’ (*SPYLab*, 2 July 2024) https://spylab.ai/blog/training-data-extraction/

Rao D, ‘The NYT vs. OpenAI Copyright Case: A Technical Introduction - Part I’ (*AI Research & Strategy*, 8 January 2024) https://deliprao.substack.com/p/the-nyt-vs-openai-copyright-case

Sakarvadia M and others, 'Mitigating Memorization In Language Models' (arXiv, 2024) arXiv:2410.02159 https://doi.org/10.48550/arXiv.2410.02159

Stober S and Dornis TW, ‘Generative AI Training and Copyright Law’ (arXiv, 2025) arXiv:2502.15858 https://doi.org/10.48550/arXiv.2502.15858

UNESCO, ‘A new expert report explores how AI is transforming culture’ (*UNESCO*, 25 November 2025) https://www.unesco.org/en/articles/new-expert-report-explores-how-ai-transforming-culture

Zhang C and others, ‘A Systematic Review on Long-Tailed Learning’ (arXiv, 2024) ArXiv:2408.00483 https://doi.org/10.48550/arXiv.2408.00483

Zhang C and others, ‘Understanding Deep Learning Requires Rethinking Generalization’ (ICLR, April 2017) https://doi.org/10.48550/arXiv.1611.03530

Ziegler A, ‘GitHub Copilot research recitation’ (*Github Blog*, 30 June 2021) https://github.blog/ai-and-ml/github-copilot/github-copilot-research-recitation/